Joseph Alexandre Heng

Neural Mechanisms Underlying Cognitive Limitations

Diss. ETH No. 29517

DISS. ETH NO. 29517

NEURAL MECHANISMS UNDERLYING COGNITIVE LIMITATIONS

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES
(Dr. sc. ETH Zurich)

presented by

JOSEPH ALEXANDRE HENG
Master of Science, EPF Lausanne

born on 22 March 1993
citizen of France and the United States of America

accepted on the recommendation of

Prof. Dr. Rafael Polanía, examiner
Prof. Dr. Todd Hare, co-examiner
Prof. Dr. Daniel Kiper, co-examiner

2023

ABSTRACT

Our brains are finite. This limitation entails that human decision-making cannot be perfect. However, it may be optimal given its restrictions. In this thesis, we consider resource limitations to study the brain and its behavior. In the first project, we study how humans perceives the number of elements in a set. In particular, we develop a model based on efficient coding given limited resources. Based on this model, we investigate which is the computational goal that underlies human perception. We find that human behavior is best captured by a model that not only maximizes accuracy, but also economizes on resources to represent the environment. In the second project, we investigate how humans deal with time limitations in perception. In particular, we investigate how humans estimate the number of elements in a set under limited time. We find that a parsimonious model based on sequential-encoding and Bayesian-decoding reproduces the variabilities and biases of human behavior. This model better captures human behavior than a thermodynamically inspired model of information processing constraints. In the third project, we study a way in which the brain deals with representing a stimuli rich world with finite resources by studying non-spatial attention. In particular, we study the role of fluctuations in excitability states in the prefrontal cortex. With a combination of neuroimaging and neurostimulation, we find these fluctuations to be causally involved in the top-down control of non-spatial attention. Finally, in the last project, we study a disorder related to decision-making by studying obesity. In particular, we investigate how dietary decision-making differs between individuals with and without obesity. We find no difference in between groups when participants rate their willingness to eat different food items. However, we observe differences in the influence of different nutritional and non-nutritional attributes as well as overt attention on food choices Altogether, this work shows the relevance of considering resource limitations to understand behavior and the brain.

RÉSUMÉ

Nos cerveaux sont limités. Cette limitation implique que la prise de décision humaine ne peut être parfaite. Cependant, elle peut être optimale compte tenu de ses restrictions. Dans cette thèse, nous considérons les limitations de ressources pour étudier le cerveau et son comportement. Dans le premier projet, nous étudions comment les humains perçoivent le nombre d'éléments dans un ensemble. En particulier, nous développons un modèle basé sur un codage efficace compte tenu des ressources limitées. Sur la base de ce modèle, nous cherchons à déterminer quel est l'objectif informationnel qui sous-tend la perception humaine. Nous constatons que le comportement humain est mieux saisi par un modèle qui non seulement maximise la précision, mais aussi économise les ressources pour représenter l'environnement. Dans le second projet, nous étudions la manière dont les humains prennent en compte les limitations temporelles dans la perception. En particulier, nous étudions comment les humains estiment le nombre d'éléments d'un ensemble dans un temps limité. Nous constatons qu'un modèle parcimonieux basé sur l'encodage séquentiel et le décodage bayésien reproduit les variabilités et les biais du comportement humain. Ce modèle rend mieux compte du comportement humain qu'un modèle d'inspiration thermodynamique des contraintes du traitement de l'information. Dans le troisième projet, nous étudions la manière dont le cerveau représente un monde riche en stimuli avec des ressources limitées en étudiant l'attention non spatiale. En particulier, nous étudions le rôle des fluctuations des états d'excitabilité dans le cortex préfrontal. Grâce à une combinaison de neuroimagerie et de neurostimulation, nous constatons que ces fluctuations sont impliquées de manière causale dans le contrôle descendant de l'attention non spatiale. Enfin, dans le dernier projet, nous étudions un trouble lié à la prise de décision en étudiant l'obésité. En particulier, nous étudions comment la prise de décision alimentaire diffère entre les individus avec et sans obésité. Nous ne constatons aucune différence entre les groupes lorsque les participants évaluent leur volonté de manger différents aliments. Cependant, nous observons des différences dans l'influence de différents attributs nutritionnels et non nutritionnels ainsi que de l'attention manifeste sur les choix alimentaires. Dans l'ensemble, ce travail montre la pertinence de prendre en compte les limitations de ressources pour comprendre le comportement et le cerveau.

ACKNOWLEDGEMENTS

I had the chance to pursue my PhD at the Decision Neuroscience Lab. If I had to express my most important lesson on decision-making, it would be that surrounding yourself with good people is always a good decision. Success cannot simply be explained by the work of an individual, as if he were isolated from the world, but will ultimately also depend on the support this individual has received. This support has been fundamental in the development of this work.

First, I would like to thank my supervisor Prof. Dr. Rafael Polanía, for being an invaluable scientific mentor. I believe that I will never have a scientific conversation with Rafa where I will not learn something new or to think about a problem is a new way. Rafa has given me the opportunities to be curious and try new approaches while always being available to answer my questions and pull me back in the known path when necessary. Rafa has shown me that leadership can be done through patience and support, and has created an exceptionally positive work environment.

I would also like to thank my co-supervisors, Prof. Dr. Todd Hare and Prof. Dr. Daniel Kiper for their valuable feedback and evaluating my work.

I would like to thank all the past and present members of the (extended) Decision Neuroscience Lab whom I had the chance of working (among other activities) alongside: Fabian, Giovanna, Jeroen, Valeriia, Silvia, Tena, Simon, Iurii, Giuseppe, Stephi, Manu, Caroline, Giulia, Gabriela and Reza.

I would like to thank my collaborators: Dr. Chloé Joray, Prof. Dr. Lia Bally, Johannes Burkard, Lucas Kohler, Dr. Cédric Sax, Prof. Dr. Erich Windhab and Prof. Dr. Michael Woodford. These collaborations have deeply enriched my thoughts and my work.

I would like to thank members for the Neural Control of Movement Lab and the Zurich Center for Neuroeconomics whom I had the chance to interact with. They have made Zurich a more scientifically dynamic and socially fun city. A particular thanks to Sanne and Marcus for their tips with fMRI analysis.

I would like to thank Dr. Roger Lüchinger, Samuel Stettler and Dr. Daniel Woolley for their technical support, as well as Sonja Bamert, Nicole Hintermeister, Dr. Maria Willecke and Dr. Xue Zhang for their administrative support.

I would like to thank the students whom I had the chance of supervising. I hope I have taught them as much as they have taught me.

I would like to thank the many participants for participating in our studies. This work could not have been done without their participation.

I would like to thank my friends for their supporting me throughout my thesis (in particular to not work on my thesis). A special thanks to Isa, Lilly, Tom and Niko for hosting me when needed and James for teaching me to send it.

I would like to thank my family for their love and support which has never dwindled. Without it, I would not have had the opportunity to carry out a PhD.

Finally I would like to thank Claire for supporting me (et de m'avoir supporté), for her patience and for sharing her adventures with me. Choosing to share this experience with you was the best decision of my life.

CONTENTS

1

INTRODUCTION

1.1. General introduction

It is estimated that people read at a speed of 238 words per minute [1]. As this manuscript contains 49 284 words, it will take 3 hours and 45 minutes for the average reader to finish it. The reader may prefer to spend their time on other activities. Therefore, as a writer I must limit the size of my thesis. This leads to the following problem. How can I communicate the most information while limiting the amount of text?

Fig 1.1. shows the numbers of occurrences of words in this manuscript depending on their length. Notice that, with the exception of one and two letter words, shorter words are used more often than longer words. This relation has been found in many languages [2] and is known as Zipf's law [3]. Although there is a debate about the origins of this law [4], one noticeable explanation is efficiency [5]. By assigning shorter words to meanings that are often used and longer words to meanings that are rarely used, we cause shorter words to appear more often. This in turn reduces the length of texts (and the number of syllables) and allows us to communicate more information for a given length of text. In other words, we can do more with a limited amount of resources. Throughout this thesis, this concept will be central and we will refer to it as efficiency[1].

The brain is our decision-making organ. Although it only weighs about 2% of our weight, it consumes 20% of our energy [6]. This means that if our ancestors brains could make better decisions while using less energy (i.e., if their brains were more efficient), they would experience a significant increase in their energy budget which would likely increase their chance of

1 Here we define efficiency as doing more with a limited amount of resources compared to a random or constant approach. Another definition of efficiency is doing the best possible given the limited amount of resources. We will refer to this optimum as peak efficiency or optimal efficiency.

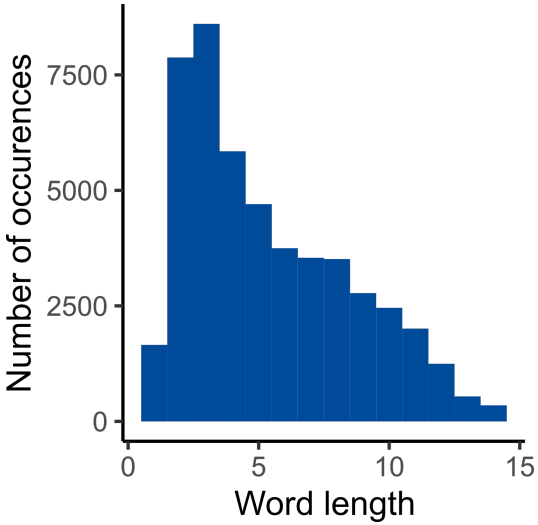


Figure 1.1: **Occurrence of words in this manuscript depending on their length.** With the exception of one and two letter words, shorter words occur more often than longer words. This relation is known as Zipf's law [3].

survival [7][2]. Following this reasoning, we can assume that our brains (and our decision-making) are shaped by the principle of efficiency.

The efficiency of the brain can be first considered on a temporal scale. The alertness of the brain varies over time. These changes in arousal are mediated in part by the locus coeruleus (LC) and its noradrenalergic projections across the brain [9]. Increases in LC activity (and thus arousal) correlate with pupil dilation in isoluminant environments [10, 11]. When we are faced with a difficult task, our pupil size (and thus arousal) increases [12–14]. This effect is also found when we are in a stressful situation [15] and causes an increase in neural metabolic demand [16]. Conversely, the LC activity decreases during sleep [17], which is associated with a decrease in metabolic activity [18, 19]. Financial incentives can provide additional evidence on the variability of brain activity on a temporal scale. Participants have higher performance on a perception task when they are financially incentivized to respond accurately [20], which has been related to their pupil size [21]. Altogether, these reports indicate that the activity of the brain is not constant,

2 For simplicity we only mention metabolic limitations. Other limitations such as constraints on wiring can be considered [8], however the conclusion remains the same: there is an evolutionary pressure for the brain to be efficient

but rather varies depending on needs of the individual. Therefore, the brain decides when to spend more or less energy depending on the context: the *when* of brain activity is efficient.

Secondly, we can consider what information is processed by the brain. We live in environments with much more information than we could ever process. Therefore, the brain must decide which information to process and which to ignore. One of the mechanisms the brain uses to solve this problem is attention. Attention consist in processing relevant information while ignoring irrelevant information. Attention has been found across modalities, however it has been mostly studies in the visual modality. Visual attention studies typically make distinctions between overt and covert attention, and distinguish between spatial and non-spatial attention [22]. Attention can be oriented by top-down goal directed intentions, for instance to focus on the road while cycling, and by bottom-up salient signals, for instance noticing an ice-cream shop on the way. Rather than processing all stimuli equally, the brain decides which computations are worth processing and which are not: the *what* of brain activity is efficient [3].

Finally, we may consider *how* the brain processes information. One aspect to consider is the code that the brain uses to represent information. Any neural network will have a finite amount of noticeably distinct states. How should the infinitely different stimulus states be mapped to a finite amount of neural states? Considering the principles of efficiency, one can respond that the stimulus states that occur more frequently should be assigned more neural states than infrequent stimulus states (similarly to Zipf's law). This principle is known as efficient coding [24][4]. It has been observed that low-level perceptual systems follow this principle [25], for example orientation perception. Cardinal orientations (i.e., horizontal and vertical) occur more frequently in our environment than non-cardinal orientations. The tuning curves of macaque V1 cells are tuned as predicted by efficient coding: more cells have preferred firing rates for orientations that occur more commonly [25]. This is consistent with the "oblique effect" observed in behavior: people perceive better horizontal and vertical orientations [26]. Interestingly, deep

3 Here I claim that attention is more efficient than considering all information equally. I do not claim that attention achieves peak efficiency. In fact, it has recently been suggested that the information gathering behavior of humans is suboptimal [23]. It is currently unknown if a resource-rational model can explain these suboptimalities.

4 Efficient coding is a form of peak efficiency, as it can be proven that it is the most efficient way of representing a stimulus. I do not make the same claim about arousal and attention, which I simply consider as more efficient than a constant neural activity in the case of arousal, or an equal consideration of all stimuli in the case of attention.

neural networks trained on image recognition tasks [27] or on orientation discrimination tasks [28] also exhibit this oblique effect, underlying the importance of efficient coding.

Another aspect to consider is neural dynamics. It has been observed that neural activity is rhythmic [29]. The frequency of this rhythmic neural activity can vary and is typically studied between 1 and 100Hz. Researchers have identified common frequency bands found in the brain. It has been suggested that high frequencies correspond to local neural computations and lower frequencies correspond to long range coordination [30]. Notably, these frequencies bands are mostly conserved across species even if their brain sizes are different [31]. The origins of these oscillations is subject to a long debate: are they a mechanism of neural communication or simply a by-product of neural computations [32]. Recently, it has been suggested that these oscillations are consistent with efficient coding. Although it is counterintuitive, as the principle of efficient coding favors reducing redundancy and thus temporal correlations, Chalk et al. [33] found that when taking into account synaptic delays, a network based on efficient coding has synchronous activity. Interestingly, the performance of the network increases when an optimal amount of noise is added (the idea that noise can be beneficial will be discussed in chapter 2). These results suggest that neural oscillations may be the result of efficient coding.

Considering how the brain processes information has been instrumental in economics. Behavioral economists were puzzled by seemingly irrational behavior. For instance, individuals preferring option A to B and B to C could also prefer C to A. This violation of intransitivity contradicts traditional economic models such as expected utility theory [34]. Simon [35, 36] proposed the idea of "bounded rationality", according to which individuals are limited in their capacities to process information and thus to make rational decisions. This notion has been central to develop models that can explain seemingly irrational behavior, in particular by assuming that the decision maker makes efficient use of their limited capacity [37].

In this thesis, we will study the brain and decision-making by considering the efficiency of the brain. In chapters 2 and 3, we will investigate how the brain encodes numerosity by considering the *how* of neural efficiency. In chapter 4, we will study the *what* of neural efficiency by investigating the neural mechanisms of non-spatial attention. Finally, in chapter 5, we will consider the *what* of neural efficiency to study a decision-making disorder by

investigating the effect of attention in dietary decision-making and how it influences differently individual with and without obesity.

The overarching goal of this thesis is to further our knowledge of brain and behavior to better treat, diagnose and prevent behavior disorders, inspire advances in artificial intelligence [38–40] and expand our knowledge of what we are.

1.2. Chapter overview

1.2.1 Chapter 2

In chapter 2, we use the *how* of neural efficiency to study the representation of numerosity. Humans (and other animals, including monkeys [41], rats [42], pigeons [43], zebrafish [44] and bees [45]) have the ability to approximately estimate the number of elements in a stimulus set without counting (e.g., dots on a screen or tones in an audio recording)[5]. This approximate number system (ANS) obeys Weber's law: the accuracy of discrimination between two stimuli is proportional to the ratio between the two stimuli [47] (a comment on the mechanisms of Weber's law is presented in Appendix A). It has been suggested that the existence of the ANS across species suggest it provides a strong benefit for survival [48]. Many decisions involve the notion of numerosity, for instance when foraging (one must decide which areas contains more resources) and social interaction (one must know which group is more numerous to predict which one would win in a fight). Notice that these decisions can also provide an intuition of why the ANS follows Weber's law. Fighting two instead of one adversary makes a bigger difference than fighting 31 instead of 30 adversaries, even though the difference in adversaries is only one in both cases.

Our study formally investigates if the ANS is based on an efficient neural code. Assuming a simple neural network, we derive formal rules that the neural firing probabilities must follow in order to be efficient. These encoding rules depend on the environmental distribution of numerosities and, importantly, the goal of the decision maker. These goals could be maximizing accuracy (i.e., maximizing the number of correct discriminations in a binary choice task), maximizing expected value (i.e., maximizing the number of elements chosen in

5 This ability has also been observed in deep neural networks trained for image recognition. In particular, researchers observed artificial neurons that responded to a preferred numerosity [46].

a binary choice task) or maximizing accuracy while economizing the resources used to represent the environmental distribution of numerosities.

Our results indicate that the human ANS is best described by an efficient coding scheme that follows the goal of maximizing accuracy while economizing the number of resources used to represent the prior distribution. In addition, we found that the human ANS adapts to new environments, which is consistent as the goal of economizing resources to represent the prior allows for a rapid adaptation. Importantly, this efficient coding model performs better than logarithmic encoding [49], a previously proposed descriptive model of the ANS.

1.2.2 Chapter 3

Similar to chapter 2, we also investigate the ANS in chapter 3. However, instead of studying the ANS in a discrimination task, we focus on an estimation task (i.e., reporting the number of dots seen on a screen). Previous work investigating the numerosity estimates of participants have found that their responses are variable and biased. We investigate if the variability and biases in numerosity estimates can be explained by the *how* of neural efficiency, by considering that the brain makes an efficient use of its limited resources. As in chapter 2, we consider numerosity estimation as a two stage process. First, a numerosity is encoded with noise into a representation. The longer the exposure time of the stimulus, the lower the encoding noise. Then, the decision-maker decodes this representation with Bayesian inference taking into account the statistics of the environment. This parsimonious model accurately predicts biases and variances observed in the human ANS. We fit this model to human behavior data and compare the fits to a recently proposed model that also consider resource limitation but using a framework inspire by thermodynamics [50]. We find that our model better captures the human ANS.

1.2.3 Chapter 4

In chapter 4, we investigate the *what* of the efficiency of the brain by studying the mechanisms of non-spatial attention. Non-spatial attention refers to the ability to enhance the representation of specific features (or objects). For instance, when foraging fruit one may enhance their processing of color to find even the hardest to find fruit, but may choose to enhance their processing of motion when they suspect a predator is hiding behind a bush. In both these examples, the individual decides which features to attend to: there is a

top-down control of attention. Baldauf and Desimone found that the inferior frontal junction (IFJ) plays a role in the top-down control of attention [51]. Participants were presented a stream of overlapping faces and houses that came in and out of visibility at different frequencies. When participants were instructed to attend to faces, the IFJ was tagged to the presentation frequency of the faces. However, when attending to houses, the IFJ's activity followed the presentation frequency of houses. In addition, they observed an increase in gamma synchrony between the IFJ and the sensory areas. These results can be added to a list of finding that have observed neural synchronization during attention task [52, 53]. The presence of neural synchronization in cognitive tasks including non-spatial attention tasks has lead to the development of the "communication through coherence" hypothesis [54], which states that neural oscillations are synchronized fluctuations in excitability states which provides windows of communication during which neurons can communicate and windows during which communication will fail. Therefore, neural oscillations are used by the brain to coordinate neural activity. This contrasts with the view that neural oscillations are simply a by-product of neural communication. To test these hypothesis, we must manipulate the levels of neural synchronization, which can be done with non-invasive brain stimulation [55, 56]. In our study, we test the causal role of neural oscillations in non-spatial attention. We develop a paradigm in which images of an indoor or outdoor scene are presented with overlapping moving dots. The participants must either attend to the content of the scene or to the direction of the dots. Importantly, the stimuli are presented as a series coming in and out of visibility (controlled by phase shuffling the images) with a period of 700ms. Our fMRI results indicate that the IFJ is involved in this non-spatial attention task, as well as the parahippocampal place area (PPA) when the participants attend to scenes and the middle temporal visual area (MT) when they attend to motion. Our EEG results show that both frontal and occipital regions are tagged by the presentation frequency and that there is a delay between the activity in the IFJ and the sensory areas. Based on these results, we non-invasively stimulate the IFJ using transcranial alternating current stimulation (tACS), which has been shown to manipulate the cortical excitability levels in an alternating manner, similar to a neural oscillation. We use two stimulation conditions, either "in-phase" with the presentation stimulus, using a delay 95ms based on our EEG results, or "out-of-phase" with the presentation stimulus, using a delay of 445ms. We hypothesized that if the synchrony of excitability states between the IFJ and sensory areas is a mechanism of the top-down control of non-spatial attention, the "in-phase"

condition would enhance performance and the "out-of-phase" condition would decrease it. However, if these oscillations are not important, both conditions would yield similar performance. We find that the performance on the task when participants are asked to attend to the direction of the dots is higher in the "in-phase" stimulation than the "out-of-phase" condition, thus showing a mechanistic role of synchrony of excitability states between the IFJ and sensory areas in non-spatial attention.

1.2.4 Chapter 5

In chapter 5, we apply our efficiency framework to study a disorder. We focus on the *what* of neural efficiency, by studying how attention influences dietary decision-making differently in individuals with and without obesity. Obesity poses a large health [57], economic [58] and environmental [59] burden. It is well know that dietary decision-making is different in obesity. For example, individuals with obesity tend to eat larger meals than individuals without obesity [60]. However, the role of attention in dietary decision-making in obesity has not been studied.

Dietary decision-making is fundamentally multi-attribute. In order to consider the choice between and apple and a cake, one may consider different nutritional attributes (e.g., how much sugar is in the apple and the cake) and non-nutritional attributes (e.g., the colorfulness of the apple and cake). These attributes must be combined to construct a representation of value (or be compared at the attribute level [61]). These attributes may be weighted differently depending on the current goal of the decision maker [62]. This weighting of attributes is reminiscent of the concept of non-spatial attention discussed in the perceptual domain, where specific features are enhanced to influence perception. FRMI studies have identifies areas of the brain involved in influencing the weighting of attributes during valuation [63]. Although these areas are distinct from the IFJ responsible for non-spatial attention, these areas are also located in the dorsolateral prefrontal cortex, which suggest weighting of attributes in the perceptual and valuation domain, although distinct, may share some similarities.

Another aspect of attention in the valuation domain is overt attention. When faced with a binary choice, people tend to choose the option that they looked at longer [64, 65]. This suggests that the brain does not have the resources to simultaneously process both options in an equal manner and is therefore biased to choose the option that is considered more.

We study how different nutritional and non-nutritional attributes and gaze influence dietary decision-making, and how these influences differ in obesity. We find that dietary decision-making is influenced by many nutritional and non-nutritional attributes, as well as the gaze of the participants. We find no significant difference when participants are asked to rate their willingness to eat different food items. However, we do find significant differences when participants are asked to make a choice between different food items. These differences allow us to accurately classify the BMI group of the participants. Therefore, we find that the influence of different nutritional and non-nutritional attributes and overt attention is different in individuals with obesity. Further investigating these results could lead to a better treatment and prevention of obesity.

2

EFFICIENT SAMPLING AND NOISY DECISIONS

J. A. Heng, M. Woodford, R. Polanía Efficient sampling and noisy decisions. eLife (2020).
doi:10.7554/eLife.54962

Contributions

Conceptualization, Formal analysis, Data analysis, Investigation, Visualization, Writing.

2.1. Abstract

Human decisions are based on finite information, which makes them inherently imprecise. But what determines the degree of such imprecision? Here, we develop an efficient coding framework for higher-level cognitive processes in which information is represented by a finite number of discrete samples. We characterize the sampling process that maximizes perceptual accuracy or fitness under the often-adopted assumption that full adaptation to an environmental distribution is possible, and show how the optimal process differs when detailed information about the current contextual distribution is costly. We tested this theory on a numerosity discrimination task, and found that humans efficiently adapt to contextual distributions, but in the way predicted by the model in which people must economize on environmental information. Thus, understanding decision behavior requires that we account for biological restrictions on information coding, challenging the often-adopted assumption of precise prior knowledge in higher-level decision systems.

2.2. Introduction

'We rarely know the statistics of the messages completely, and our knowledge may change ... what is redundant today was not necessarily redundant yesterday.' – Horace Barlow [66].

It has been suggested that the rules guiding behavior are not arbitrary, but follow fundamental principles of acquiring information from environmental regularities in order to make the best decisions. Moreover, these principles should incorporate strategies of information coding in ways that minimize the costs of inaccurate decisions given biological constraints on information acquisition, an idea known as efficient coding [24, 67–69]. While early applications of efficient coding theory have primarily been to early stages of sensory processing [70–72], it is worth considering whether similar principles may also shape the structure of internal representations of higher-level concepts, such as the perceptions of value that underlie economic decision making [73–75]. In this work, we contribute to the efficient coding framework applied to cognition and behavior in several respects.

A first aspect concerns the range of possible internal representation schemes that should be considered feasible, which determines the way in which greater precision of discrimination in one part of the stimulus space requires less precision of discrimination elsewhere. Implementational architectures proposed in previous work assume a population coding scheme in which different neurons have distinct 'preferred' stimuli [71, 72]. While this is clearly relevant for some kinds of low-level sensory features such as orientation, it is not obvious that this kind of internal representation is used in representing higher-level concepts such as economic values. We instead develop an efficient coding theory for a case in which an extensive magnitude (something that can be described by a larger or smaller number) is represented by a set of processing units that 'vote' in favor of the magnitude being larger rather than small. The internal representation therefore necessarily consists of a finite collection of binary signals.

Our restriction to representations made up of binary signals is in conformity with the observation that neural systems at many levels appear to transmit information via discrete stochastic events [76, 77]. Moreover, cognitive models with this general structure have been argued to be relevant for higher-order decision problems such as value-based choice. For example, it has been suggested that the perceived values of choice options are constructed by acquiring samples of evidence from memory regarding the emotions evoked

by the presented items [78]. Related accounts suggest that when a choice must be made between alternative options, information is acquired via discrete samples of information that can be represented as binary responses (e.g., 'yes/no' responses to queries) [79, 80]. The seminal decision by sampling (DbS) theory [81] similarly posits an internal representation of magnitudes relevant to a decision problem by tallies of the outcomes of a set of binary comparisons between the current magnitude and alternative values sampled from memory. The architecture that we assume for imprecise internal representations has the general structure of proposals of these kinds; but we go beyond the above-mentioned investigations, in analyzing what an efficient coding scheme consistent with our general architecture would be like.

A second aspect concerns the objective for which the encoding system is assumed to be optimized. Information maximization theories [70–72] assume that the objective should be maximal mutual information between the true stimulus magnitude and the internal representation. While this may be a reasonable assumption in the case of early sensory processing, it is less obvious in the case of circuits involved more directly in decision making, and in the latter case an obvious alternative is to ask what kind of encoding scheme will best serve to allow accurate decisions to be made. In the theory that we develop here, our primary concern is with encoding schemes that maximize a subject's probability of giving a correct response to a binary decision. However, we compare the coding rule that would be optimal from this standpoint to one that would maximize mutual information, or to one that would maximize the expected value of the chosen item.

Third, we extend our theory of efficient coding to consider not merely the nature of an efficient coding system for a single environmental frequency distribution assumed to be permanently relevant —so that there has been ample time for the encoding rule to be optimally adapted to that distribution of stimulus magnitudes —but also an efficient approach to adjusting the encoding as the environmental frequency distribution changes. Prior discussions of efficient coding have often considered the optimal choice of an encoding rule for a single environmental frequency distribution that is assumed to represent a permanent feature of the natural environment [70, 71]. Such an approach may make sense for a theory of neural coding in cortical regions involved in early-stage processing of sensory stimuli, but is less obviously appropriate for a theory of the processing of higher-level concepts such as economic value, where the idea that there is a single permanently relevant frequency distribution of magnitudes that may be encountered is doubtful.

A key goal of our work is to test the relevance of these different possible models of efficient coding in the case of numerosity discrimination. Judgments of the comparative numerosity of two visual displays provide a test case of particular interest given our objectives. On the one hand, a long literature has argued that imprecision in numerosity judgments has a similar structure to psychophysical phenomena in many low-level sensory domains [82, 83]. This makes it reasonable to ask whether efficient coding principles may also be relevant in this domain. At the same time, numerosity is plainly a more abstract feature of visual arrays than low-level properties such as local luminosity, contrast, or orientation, and therefore can be computed only at a later stage of processing. Moreover, processing of numerical magnitudes is a crucial element of many higher-level cognitive processes, such as economic decision making; and it is arguable that many rapid or intuitive judgments about numerical quantities, even when numbers are presented symbolically, are based on an 'approximate number system' of the same kind as is used in judgments of the numerosity of visual displays [82, 84]. It has further been argued that imprecision in the internal representation of numerical magnitudes may underly imprecision and biases in economic decisions [85, 86].

It is well-known that the precision of discrimination between nearby numbers of items decreases in the case of larger numerosities, in approximately the way predicted by Weber's Law, and this is often argued to support a model of imprecise coding based on a logarithmic transformation of the true number [82, 83]. However, while the precision of internal representations of numerical magnitudes is arguably of great evolutionary relevance [48, 87], it is unclear why a specifically logarithmic transformation of number information should be of adaptive value, and also whether the same transformation is used independent of context [47, 88]. Here, we report new experimental data on numerosity discrimination by human participants, where we find that our data are most consistent with an efficient coding theory for which the performance measure is the frequency of correct comparative judgments, and where people economize on the costs associated to learn about the statistics of the environment.

2.3. Results

A general efficient sampling framework

We consider a situation in which the objective magnitude of a stimulus with respect to some feature can be represented by a quantity v. When the stimulus is presented to an observer, it gives rise to an imprecise representation r in the nervous system, on the basis of which the observer produces any required response. The internal representation r can be stochastic, with given values being produced with conditional probabilities $p(r|v)$ that depend on the true magnitude. Here, we are more specifically concerned with discrimination experiments, in which two stimulus magnitudes v_1 and v_2 are presented, and the subject must choose which of the two is greater. We suppose that each magnitude v_i has an internal representation r_i, drawn independently from a distribution $p(r_i|v_i)$ that depends only on the true magnitude of that individual stimulus. The observer's choice must be based on a comparison of r_1 with r_2.

One way in which the cognitive resources recruited to make accurate discriminations may be limited is in the variety of distinct internal representations that are possible. When the complexity of feasible internal representations is limited, there will necessarily be errors in the identification of the greater stimulus magnitude in some cases, even assuming an optimal decoding rule for choosing the larger stimulus on the basis of r_1 and r_2. One can then consider alternative encoding rules for mapping objective stimulus magnitudes to feasible internal representations. The answer to this efficient coding problem generally depends on the prior distribution $f(v)$ from which the different stimulus magnitudes v_i are drawn. The resources required for more precise internal representations of individual stimuli may be economized with respect to either or both of two distinct cognitive costs. The first goal of this work is to distinguish between these two types of efficiency concerns.

One question that we can ask is whether the observed behavioral responses are consistent with the hypothesis that the conditional probabilities $p(r|v)$ are well-adapted to the particular frequency distribution of stimuli used in the experiment, suggesting an efficient allocation of the limited encoding neural resources. The assumption of full adaptation is typically adopted in efficient coding formulations of early sensory systems [70, 89], and also more recently in applications of efficient coding theories in value-based decisions [73–75].

There is also a second cost in which it may be important to economize on cognitive resources. An efficient coding scheme in the sense described above economizes on the resources used to represent each individual new stimulus that is encountered; however, the encoding and decoding rules are assumed to be precisely optimized for the specific distribution $f(v)$ of stimuli that characterizes the experimental situation. In practice, it will be necessary for a decision maker to learn about this distribution in order to encode and decode individual stimuli in an efficient way, on the basis of experience with a given context. In this case, the relevant design problem should not be conceived as choosing conditional probabilities $p(r|v)$ once and for all, with knowledge of the prior distribution $f(v)$ from which v will be drawn. Instead, it should be to choose a rule that specifies how the probabilities $p(r|v)$ should adapt to the distribution of stimuli that have been encountered in a given context. It then becomes possible to consider how well a given learning rule economizes on the degree of information about the distribution of magnitudes associated with one's current context that is required for a given level of average performance across contexts. This issue is important not only to reduce the cognitive resources required to implement the rule in a given context (by not having to store or access so detailed a description of the prior distribution), but in order to allow faster adaptation to a new context when the statistics of the environment can change unpredictably [90].

Coding architecture

We now make the contrast between these two types of efficiency more concrete by considering a specific architecture for internal representations of sensory magnitudes. We suppose that the representation r_i of a given stimulus will consist of the output of a finite collection of n processing units, each of which has only two possible output states ('high' or 'low' readings), as in the case of a simple perceptron. The probability that each of the units will be in one output state or the other can depend on the stimulus v_i that is presented. We further restrict the complexity of feasible encoding rules by supposing that the probability of a given unit being in the 'high' state must be given by some function $\theta(v_i)$ that is the same for each of the individual units, rather than allowing the different units to coordinate in jointly representing the situation in some more complex way. We argue that the existence of multiple units operating in parallel effectively allows multiple repetitions of the same 'experiment', but does not increase the complexity of the kind of test that can be performed. Note that we do not assume any unavoidable degree of stochasticity in the functioning of the individual units; it turns out that

in our theory, it will be efficient for the units to be stochastic, but we do not assume that precise, deterministic functioning would be infeasible. Our resource limits are instead on the number of available units, the degree of differentiation of their output states, and the degree to which it is possible to differentiate the roles of distinct units.

Given such a mechanism, the internal representation r_i of the magnitude of an individual stimulus v_i will be given by the collection of output states of the n processing units. A specification of the function $\theta(v)$ then implies conditional probabilities for each of the $2n$ possible representations. Given our assumption of a symmetrical and parallel process, the number k_i of units in the 'high' state will be a sufficient statistic, containing all of the information about the true magnitude v_i that can be extracted from the internal representation. An optimal decoding rule will therefore be a function only of k_i, and we can equivalently treat k_i (an integer between 0 and n) as the internal representation of the quantity v_i. The conditional probabilities of different internal representations are then

$$p(k_i|v_i) = \binom{n}{k}\theta(v_i)^{k_i}(1-\theta(v_i))^{n-k_i} \tag{2.1}$$

The efficient coding problem for a given environment, specified by a particular prior distribution $f(v)$, will be to choose the encoding rule $\theta(v)$ so as to allow an overall distribution of responses across trials that will be as accurate as possible (according to criteria that we will elaborate further below). We can further suppose that each of the individual processing units is a threshold unit, that produces a 'high' reading if and only if the value $v_i - \eta_i$ exceeds some threshold τ, where η_i is a random term drawn independently on each trial from some distribution $f\eta$ (Figure 2.1). The encoding function $\theta(v)$ can then be implemented by choosing an appropriate distribution $f\eta$. This implementation requires that $\theta(v)$ be a non-decreasing function, as we shall assume.

Limited cognitive resources

One measure of the cognitive resources required by such a system is the number n of processing units that must produce an output each time an individual stimulus v_i is evaluated. We can consider the optimal choice of $f\eta$ in order to maximize, for instance, average accuracy of responses in a given environment $f(v)$, in the case of any bound n on the number of units that can be used to represent each stimulus. But we can also consider the amount

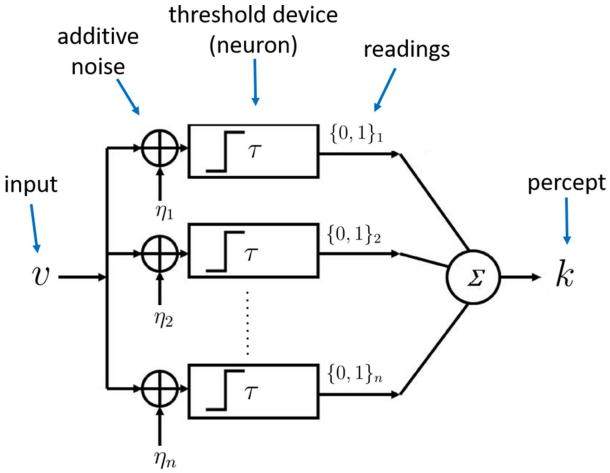


Figure 2.1: **Architecture of the sampling mechanism.** Each processing unit receives noisy versions of the input v, where the noisy signals are i.i.d. additive random signals independent of v. The output of the neuron for each sample is 'high' (one) reading if $v - \eta > \tau$ and zero otherwise. The noisy percept of the input is simply the sum of the outputs of each sample given by k.

of information about the distribution $f(v)$ that must be used in order to decide how to encode a given stimulus v_i. If the system is to be able to adapt to changing environments, it must determine the value of θ (the probability of a 'high' reading) as a function of both the current v_i and information about the distribution f, in a way that must now be understood to apply across different potential contexts. This raises the issue of how precisely the distribution f associated with the current context is represented for purposes of such a calculation. A more precise representation of the prior (allowing greater sensitivity to fine differences in priors) will presumably entail a greater resource cost or very long adaptation periods.

We can quantify the precision with which the prior f is represented by supposing that it is represented by a finite sample of m independent draws $\tilde{v_1}, \ldots, \tilde{v_m}$ from the prior (or more precisely, from the set of previously experienced values, an empirical distribution that should after sufficient experience provide a good approximation to the true distribution). We further assume that an independent sample of m previously experienced values is used by each of the processing units (Figure 2.1). Each of the n individual processing units is then in the 'high' state with probability $\theta(v_i; \tilde{v_1}, \ldots, \tilde{v_m})$. The complete internal representation of the stimulus v_i is then the collection

of n independent realizations of this binary-valued random variable. We may suppose that the resource cost of an internal representation of this kind is an increasing function of both n and m.

This allows us to consider an efficient coding meta-problem in which for any given values (n, m) the function $\theta(v_i; \tilde{v}_1, \ldots, \tilde{v}_m)$ is chosen so as to maximize some measure of average perceptual accuracy, where the average is now taken not only over the entire distribution of possible v_i occurring under a given prior $f(v)$, but over some range of different possible priors for which the adaptive coding scheme is to be optimized. We wish to consider how each of the two types of resource constraint (a finite bound on n as opposed to a finite bound on m) affects the nature of the predicted imprecision in internal representations, under the assumption of a coding scheme that is efficient in this generalized sense, and then ask whether we can tell in practice how tight each of the resource constraints appears to be.

Efficient sampling for a known prior distribution

We first consider efficient coding in the case that there is no relevant constraint on the size of m, while n instead is bounded. In this case, we can assume that each time an individual stimulus v_i must be encoded, a large enough sample of prior values is used to allow accurate recognition of the distribution $f(v)$, and the problem reduces to a choice of a function $\theta(v)$ that is optimal for each possible prior $f(v)$.

2.3.0.1 *Maximizing mutual information*

The nature of the resource-constrained problem to be optimized depends on the performance measure that we use to determine the usefulness of a given encoding scheme. A common assumption in the literature on efficient coding has been that the encoding scheme maximizes the mutual information between the true stimulus magnitude and its internal representation [71, 72, 74]. We start by characterizing the optimal $\theta(v)$ for a given prior distribution $f(v)$, according to this criterion. It can be shown that for large n, the mutual information between θ and k (hence the mutual information between v and k) is maximized if the prior distribution $\hat{f}$ over θ is Jeffreys' prior [91]

$$\hat{f}(\theta) = \frac{1}{\pi\sqrt{\theta(1-\theta)}} \tag{2.2}$$

also known as the arcsine distribution. Hence, the mapping $\theta(v)$ induces a prior distribution $\hat{f}$ over θ given by the arcsine distribution (Figure 2.2a, right panel). Based on this result, it can be shown that the optimal encoding rule $\theta(v)$ that guarantees maximization of mutual information between the random variable v and the noisy encoded percept k is given by (see Appendix 2.1)

$$\theta(v) = [sin(\frac{\pi}{2}F(v))]^2, \tag{2.3}$$

where $F(v)$ is the CDF of the prior distribution $f(v)$.

2.3.0.2 *Accuracy maximization for a known prior distribution*

So far, we have derived the optimal encoding rule to maximize mutual information. However, one may ask what the implications are of such a theory for discrimination performance. This is important to investigate given that achieving channel capacity does not necessarily imply that the goals of the organism are also optimized [92]. Independent of information maximization assumptions, here, we start from scratch and investigate what are the necessary conditions for minimizing discrimination errors given the resource-constrained problem considered here. We solve this problem for the case of two alternative forced choice tasks, where the average probability of error is given by (see Appendix 2.2)

$$E[error] = \iint P_{error}[\theta(v_1), \theta(v_2)]\hat{f}(\theta_1)\hat{f}(\theta_2)d\theta_1 d\theta_2 \tag{2.4}$$

where $P_{error}[]$ represents the probability of erroneously choosing the alternative with the lowest value v given a noisy percept k (assuming that the goal of the organism in any given trial is to choose the alternative with the highest value). Here, we want to find the density function $\hat{f}(\theta)$ that guarantees the smallest average error (Equation 4). The solution to this problem is (Appendix 2.2)

$$\hat{f}(\theta) = \frac{1}{\pi\sqrt{\theta(1-\theta)}} \tag{2.5}$$

which is exactly the same prior density function over θ that maximizes mutual information (Equation 2). Crucially, please note that we have obtained this expression based on minimizing the frequency of erroneous choices and not the

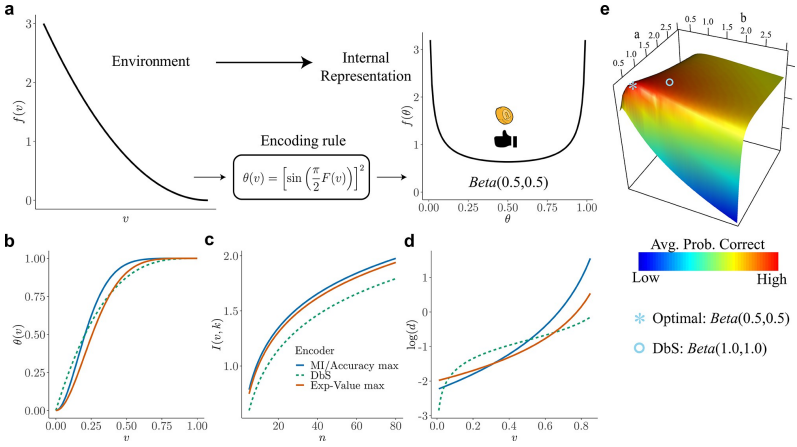


Figure 2.2: **Overview of our theory and differences in encoding rules. (a)** Schematic representation of our theory. Left: example prior distribution $f(v)$ of values v encountered in the environment. Right: Prior distribution in the encoder space (Equation 2) due to optimal encoding (Equation 3). This optimal mapping determines the probability θ of generating a 'high' or 'low' reading. The ex-ante distribution over θ that guarantees maximization of mutual information is given by the arcsine distribution (Equation 2). **(b)** Encoding rules $\theta(v)$ for different decision strategies under binary sampling coding: accuracy maximization (blue), reward maximization (red), DbS (green dashed). **(c)** Mutual information $I(v,k)$ for the different encoding rules as a function of the number of samples n. As expected $I(v,k)$ increases with n, however the rule that results in the highest loss of information is DbS. **(d)** Discriminability thresholds d (log-scaled for better visualization) for the different encoding rules as a function of the input values v for the prior $f(v)$ given in panel a. **(e)** Graphical representation of the perceptual accuracy optimization landscape. We plot the average probability of correct responses for the large-n limit using as benchmark a Beta distribution with parameters a and b. The blue star shows the average error probability assuming that $f(\theta)$ is the arcsine distribution (Equation 2), which is the optimal solution when the prior distribution f in known. The blue open circle shows the average error probability based on the encoding rule assumed in DbS, which is located near the optimal solution. Please note that when formally solving this optimization problem, we did not assume a priori that the solution is related to the beta distribution. We use the beta distribution in this figure just as a benchmark for visualization. Detailed comparison of performance for finite n samples is presented in Appendix 2.7.

maximization of mutual information as a goal in itself. This provides a further (and normative) justification for why maximizing mutual information under this coding scheme is beneficial when the goal of the agent is to minimize discrimination errors (i.e., maximize accuracy).

2.3.0.3 *Optimal noise for a known prior distribution*

Based on the coding architecture presented in Figure 2.1, the optimal encoding function $\theta(v)$ can then be implemented by choice of an appropriate distribution $f\eta$. It can be shown that discrimination performance can be optimized by finding the optimal noise distribution $f\eta$ (Appendix 2.3) [93]

$$f_\eta(v) = \frac{\pi}{2} sin[\pi(1 - F(\tau - v))] f(\tau - v) \tag{2.6}$$

Remarkably, this result is independent of the number of samples n available to encode the input variable, and generalizes to any prior distribution f (recall that F is defined as its cumulative density function).

This result reveals three important aspects of neural function and decision behavior: First, it makes explicit why a system that evolved to code information using a coding scheme of the kind assumed in our framework must be necessarily noisy. That is, we do not attribute the randomness of peoples' responses to a particular set of stimuli or decision problem to unavoidable randomness of the hardware used to process the information. Instead, the relevant constraints are assumed to be the limited set of output states for each neuron, the limited number of neurons, and the requirement that the neurons operate in parallel (so that each one's output state must be statistically independent of the others, conditional on the input stimulus). Given these constraints, we show that it is efficient for the operation of the neurons to be random. Second, it shows how the nervous system may take advantage of these noisy properties by reshaping its noise structure to optimize decision behavior. Third, it shows that the noise structure can remain unchanged irrespective of the amount of resources available to guide behavior (i.e., the noise distribution f_η does not depend on n, Equation 6). Please note however, that this minimalistic implementation does not directly imply that the samples in our algorithmic formulation are necessarily drawn in this way. We believe that this implementation provides a simple demonstration of the consequences of limited resources in systems that encode information based on discrete stochastic events [77]. Interestingly, it has been shown that this

minimalistic formulation can be extended to more realistic population coding specifications [94].

2.3.0.4 *Efficient coding and the relation between environmental priors and discrimination*

The results presented above imply that this encoding framework imposes limitations on the ability of capacity-limited systems to discriminate between different values of the encoded variables. Moreover, we have shown that error minimization in discrimination tasks implies a particular shape of the prior distribution of the encoder (Equation 5) that is exactly the prior density that maximizes mutual information between the input v and the encoded noisy readings k (Equation 2, Figure 2.2a right panel). Does this imply a relation between prior and discriminability over the space of the encoded variable? Intuitively, following the efficient coding hypothesis, the relation should be that lower discrimination thresholds should occur for ranges of stimuli that occur more frequently in the environment or context.

Recently, it was shown that using an efficiency principle for encoding sensory variables (e.g., with a heterogeneous population of noisy neurons [25]) it is possible to obtain an explicit relationship between the statistical properties of the environment and perceptual discriminability [25]. The theoretical relation states that discriminability thresholds d should be inversely proportional to the density of the prior distribution $f(v)$. Here, we investigated whether this particular relation also emerges in the efficient coding scheme that we propose in this study.

Remarkably, we obtain the following relation between discriminability thresholds, prior distribution of input variables, and the number of limited samples n (Appendix 2.4):

$$\begin{aligned} d &= \frac{1}{\sqrt{n}\pi f(v)} \\ &\propto \frac{1}{f(v)} \end{aligned} \tag{2.7}$$

Interestingly, this relationship between prior distribution and discriminability thresholds holds empirically across several sensory modalities (Appendix 2.4), thus once again demonstrating that the efficient coding framework that we propose here seems to incorporate the right kind of constraints to explain

observed perceptual phenomena as consequences of optimal allocation of finite capacity for internal representations.

2.3.0.5 *Maximizing the expected size of the selected option (fitness maximization)*

Until now, we have studied the case when the goal of the organism is to minimize the number of mistakes in discrimination tasks. However, it is important to consider the case when the goal of the organism is to maximize fitness or expected reward [95]. For example, when spending the day foraging fruit, one must make successive decisions about which tree has more fruits. Fitness depends on the number of fruit collected which is not a linear function of the number of accurate decisions, as each choice yields a different amount of fruit.

Therefore, in the case of reward maximization, we are interested in minimizing reward loss which is given by the following expression

$$E[v(chosen)] = \iint f(v_1, v_2)[P_1(\theta(v_1), \theta(v_2))v_1 + P_2(\theta(v_1), \theta(v_2))v_2]dv_1 dv_2 \tag{2.8}$$

where $P_i(\theta(v_1), \theta(v_2))$ is the probability of choosing option i when the input values are v_1 and v_2. Thus, the goal is to find the encoding rule $\theta(v)$ which guarantees that the amount of reward loss is as small as possible given our proposed coding framework.

Here we show that the optimal encoding rule $\theta(v)$ that guarantees maximization of expected value is given by

$$\theta(v) = sin\left[\frac{\pi}{2} \cdot c \int_{-\infty}^{v} f(\tilde{v})^{2/3} d\tilde{v}\right]^2, \tag{2.9}$$

where c is a normalizing constant which guarantees that the expression within the integral is a probability density function (Appendix 2.5). The first observation based on this result is that the encoding rule for maximizing fitness is different from the encoding rule that maximizes accuracy (compare Equations 3 and 9), which leads to a slight loss of information transmission (Figure 2.2c). Additionally, one can also obtain discriminability threshold predictions for this new encoding rule. Assuming a right-skewed prior distribution, which is often the case for various natural priors in the environment

(e.g., like the one shown in Figure 2.2a), we find that discriminability for small input values is lower for reward maximization compared to perceptual maximization, however this pattern inverts for higher values (Figure 2.2d). In other words, when we intend to maximize reward (given the shape of our assumed prior, Figure 2.2a), the agent should allocate more resources to higher values (compared to the perceptual case), however without completely giving up sensitivity for lower values, as these values are still encountered more often.

2.3.0.6 *Efficient sampling with costs on acquiring prior knowledge*

In the previous section, we obtained analytical solutions that approximately characterize the optimal $\theta(v)$ in the limit as n is made sufficiently large. Note however that we are always assuming that is finite, and that this constrains the accuracy of the decision maker's judgments, while m is instead unbounded and hence no constraint.

The nature of the optimal function $\theta(v_i; \tilde{v}_1, \ldots, \tilde{v}_m)$ is different, however, when m is small. We argue that this scenario is particularly relevant when full knowledge of the prior is not warranted given the costs vs benefits of learning, for instance, when the system expects contextual changes to occur often. In this case, as we will formally elaborate below, it ceases to be efficient for θ to vary only gradually as a function of v_i, rather than moving abruptly from values near zero to values near one (Appendix 2.6). In the large-m limiting case, the distributions of sample values $(\tilde{v}_1, \ldots, \tilde{v}_m)$ used by the different processing units will be nearly the same for each unit (approximating the current true distribution $f(v)$). Then if θ were to take only the values zero and one for different values of its arguments, the n units would simply produce n copies of the same output (either zero or one) for any given stimulus v_i and distribution $f(v)$. Hence only a very coarse degree of differentiation among different stimulus magnitudes would be possible. Having θ vary more gradually over the range of values of v_i in the support of $f(v)$ instead makes the representation more informative. But when m is small (e.g., because of costs vs benefits of accurately representing the prior f), this kind of arbitrary randomization in the output of individual processing units is no longer essential. There will already be considerable variation in the outputs of the different units, even when the output of each unit is a deterministic function of $(\tilde{v}_1, \ldots, \tilde{v}_m)$, owing to the variability in the sample of prior observations that is used to assess the nature of the current environment. As we will show below, this variability will already serve to

allow the collective output of the several units to differentiate between many gradations in the magnitude of v_i, rather than only being able to classify it as 'small' or 'large' (because either all units are in the 'low' or 'high' states).

2.3.0.7 *Robust optimality of decision by sampling*

Because of the way in which sampling variability in the values $(\tilde{v}_1, \ldots, \tilde{v}_m)$ used to adapt each unit's encoding rule to the current context can substitute for the arbitrary randomization represented by the noise term η_i (see Figure 2.1), a sharp reduction in the value of m need not involve a great loss in performance relative to what would be possible (for the same limit on n) if m were allowed to be unboundedly large (Appendix 2.7). As an example, consider the case in which $m = 1$, so that each unit j's output state must depend only on the value of the current stimulus v_i and one randomly selected draw $\tilde{v}_j$ from the prior distribution $f(v)$. A possible decision rule that is radically economical in this way is one that specifies that the unit will be in the 'high' state if and only if $v_i > \tilde{v}_j$. In this case, the internal representation of a stimulus v_i will be given by the number k_i out of n independent draws from the contextual distribution $f(v)$ with the property that the contextual draw is smaller than v_i, as in the model of decision by sampling (DbS) [81]. However, it remains to be determined to what degree it might be beneficial for a system to adopt such coding strategy.

In any given environment (characterized by a particular contextual distribution $f(v)$), DbS will be equivalent to an encoding process with an architecture of the kind shown in Figure 2.1, but in which the distribution $f_\eta = f(v)$ (compare to the optimal noise distribution f_η for the full prior adaptation case in Equation 6). This makes $\theta(v)$ vary endogenously depending on the contextual distribution $f(v)$. And indeed, the way that $\theta(v)$ varies with the contextual distribution under DbS is fairly similar to the way in which it would be optimal for it to vary in the absence of any cost of precisely learning and representing the contextual distribution. This result implies that $\theta(v)$ will be a monotonic transformation of a function that increases more steeply over those regions of the stimulus space where $f(v)$ is higher, regardless of the nature of the contextual distribution. We consider its performance in a given environment, from the standpoint of each of the possible performance criteria considered for the case of full prior adaptation (i.e., maximize accuracy or fitness), and show that it differs from the optimal encoding rules under any of those criteria (Figure 2.2b–d). In particular, here, we show that using the

encoding rule employed in DbS results in considerable loss of information compared to the full-prior adaptation solutions (Figure 2.2c). An additional interesting observation is that for the strategy employed in DbS, the agent appears to be more sensitive for extreme input values, at least for a wide set of skewed distributions (e.g., for the prior distribution $f(v)$ in Figure 2.2a, the discriminability thresholds are lower at the extremes of the support of $f(v)$). In other words, agents appear to be more sensitive to salience in the DbS rule. Despite these differences, here it is important to emphasize that in general for all optimization objectives, the encoding rules will be steeper for regions of the prior with higher density. However, mild changes in the steepness of the curves will be represented in significant discriminability differences between the different encoding rules across the support of the prior distribution (Figure 2.2d).

While the predictions of DbS are not exactly the same as those of efficient coding in the case of unbounded m, under any of the different objectives that we consider, our numerical results show that it can achieve performance nearly as high as that of the theoretically optimal encoding rule; hence radically reducing the value of m does not have a large cost in terms of the accuracy of the decisions that can be made using such an internal representation (Appendix 2.7 and Figure 2.2e). Under the assumption that reducing either m or n would serve to economize on scarce cognitive resources, we formally prove that it might well be most efficient to use an algorithm with a very low value of m (even $m = 1$, as assumed by DbS), while allowing n to be much larger (Appendix 2.6, Appendix 2.7).

Crucially, here, it is essential to emphasize that the above-mentioned results are derived for the case of a particular finite number of processing units n (and a corresponding finite total number of samples from the contextual distribution used to encode a given stimulus), and do not require that n must be large (Appendix 2.6, Appendix 2.7).

2.3.0.8 *Testing theories of numerosity discrimination*

Our goal now is to compare back-to-back the resource-limited coding frameworks elaborated above in a fundamental cognitive function for human behavior: numerosity perception. We designed a set of experiments that allowed us to test whether human participants would adapt their numerosity encoding system to maximize fitness or accuracy rates via full prior adaptation as usually assumed in optimal models, or whether humans employ a

'less optimal' but more efficient strategy such as DbS, or the more established logarithmic encoding model.

In Experiment 1, healthy volunteers (n = 7) took part in a two-alternative forced choice numerosity task in which each participant completed ~ 2400 trials across four consecutive days (Methods). On each trial, they were simultaneously presented with two clouds of dots and asked which one contained more dots, and were given feedback on their reward and opportunity losses on each trial (Figure 2.3a). Participants were either rewarded for their accuracy (perceptual condition, where maximizing the amount of correct responses is the optimal strategy) or the number of dots they selected (value condition, where maximizing reward is the optimal strategy). Each condition was tested for two consecutive days with the starting condition randomized across participants. Crucially, we imposed a prior distribution $f(v)$ with a right-skewed quadratic shape (Figure 2.3b), whose parametrization allowed tractable analytical solutions of the encoding rules $\theta_A(v)$, $\theta_R(v)$ and $\theta_D(v)$, that correspond to the encoding rules for Accuracy maximization, Reward maximization, and DbS, respectively (Figure 2.3e and Methods). Qualitative predictions of behavioral performance indicate that the accuracy-maximization model is the most accurate for trials with lower numerosities (the most frequent ones), whereas the reward-maximization model outperforms the others for trials with larger numerosities (trials where the difference in the number of dots in the clouds, and thus the potential reward, is the largest, Figure 2.2d and Figure 2.3f). In contrast, the DbS strategy presents markedly different performance predictions, in line with the discriminability predictions of our formal analyses (Figure 2.2c,d).

In our modelling specification, the choice structure is identical for the three different sampling models, differing only in the encoding rule $\theta(v)$ (Methods). Therefore, answering the question of which encoding rule is the most favored for each participant can be parsimoniously addressed using a latent-mixture model, where each participant uses $\theta_A(v)$, $\theta_R(v)$ or $\theta_D(v)$ to guide their decisions (Methods). Before fitting this model to the empirical data, we confirmed the validity of our model selection approach through a validation procedure using synthetic choice data (Figure 2.3d, Figure 2.3—figure supplement 1, and Methods).

After we confirmed that we can reliably differentiate between our competing encoding rules, the latent-mixture model was initially fitted to each condition (perceptual or value) using a hierarchical Bayesian approach (Methods). Surprisingly, we found that participants did not follow the accuracy or reward

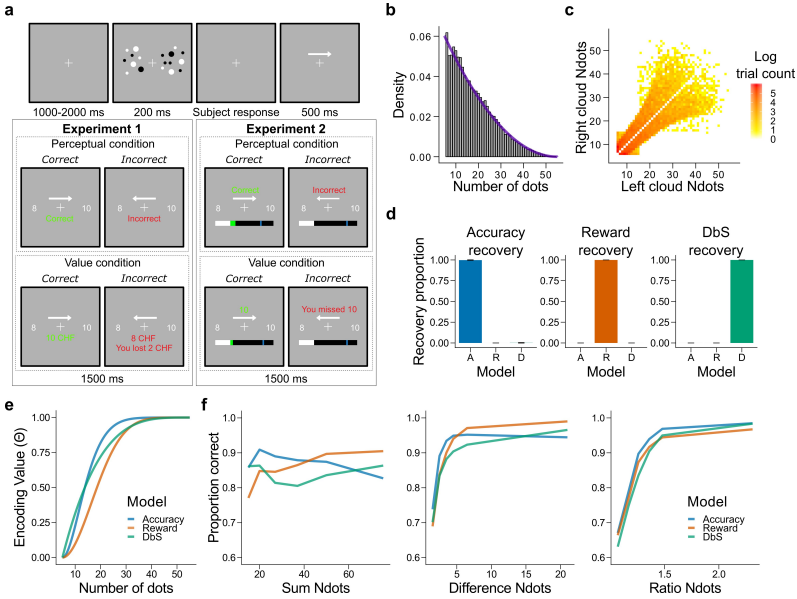


Figure 2.3: **Experimental design, model simulations and recovery. (a)** Schematic task design of Experiments 1 and 2. After a fixation period (1–2 s) participants were presented two clouds of dots (200 ms) and had to indicate which cloud contained the most dots. Participants were rewarded for being accurate (perceptual condition) or for the number of dots they selected (value condition) and were given feedback. In Experiment 2 participants collected on correctly answered trials a number of points equal to a fixed amount (perceptual condition) or a number equal to the dots in the cloud they selected (value condition) and had to reach a threshold of points on each run. (b) Empirical (grey bars) and theoretical (purple line) distribution of the number of dots in the clouds of dots presented across Experiments 1 and 2. (c) Distribution of the numerosity pairs selected per trial. (d) Synthetic data preserving the trial set statistics and number of trials per participant used in Experiment 1 was generated for each encoding rule (Accuracy (left), Reward (middle), and DbS (right)) and then the latent-mixture model was fitted to each generated dataset. The figures show that it is theoretically possible to recover each generated encoding rule. (e) Encoding function $\theta(v)$ for the different sampling strategies as a function of the input values v (i.e., the number of dots). (f) Qualitative predictions of the three models (blue: Accuracy, red: Reward, green: Decision by Sampling) on trials from Experiment 1 with $n = 25$. Performance of each model as a function of the sum of the number of dots in both clouds (left), the absolute difference between the number of dots in both clouds (middle) and the ratio of the number of dots in the most numerous cloud over the less numerous cloud (right).

optimization strategy in the respective experimental condition, but favored the DbS strategy (proportion that DbS was deemed best in the perceptual $p_{DbSfavored} = 0.86$ and value $p_{DbSfavored} = 0.93$ conditions, Figure 2.4). Importantly, this population-level result also holds at the individual level: DbS was strongly favored in 6 out of 7 participants in the perceptual condition, and seven out of seven in the value condition (Figure 2.4—figure supplement 1). These results are not likely to be affected by changes in performance over time, as performance was stable across the four consecutive days (Figure 2.4—figure supplement 2). Additionally, we investigated whether biases induced by choice history effects may have influenced our results [96–98]. Therefore, we incorporated both choice- and correctness-dependence history biases in our models and fitted the models once again (Methods). We found similar results to the history-free models ($p_{DbSfavored} = 0.87$ in perceptual and $p_{DbSfavored} = 0.93$ in value conditions, Figure 2.4c). At the individual level, DbS was again strongly favored in 6 out of 7 participants in the perceptual condition, and 7 out of 7 in the value condition (Figure 2.4—figure supplement 1).

In order to investigate further the robustness of this effect, we introduced a slight variation in the behavioral paradigm. In this new experiment (Experiment 2), participants were given points on each trial and had to reach a certain threshold in each run for it to be eligible for reward (Figure 2.3a and Methods). This class of behavioral task is thought to be in some cases more ecologically valid than trial-independent choice paradigms [99]. In this new experiment, either a fixed amount of points for a correct trial was given (perceptual condition) or an amount equal to the number of dots in the chosen cloud if the response was correct (value condition). We recruited a new set of participants (n = 6), who were tested on these two conditions, each for two consecutive days with the starting condition randomized across participants (each participant completed ∼2,560 trials). The quantitative results revealed once again that participants did not change their encoding strategy depending on the goals of the task, with DbS being strongly favored for both perceptual and value conditions ($p_{DbSfavored} = 0.999$ and $p_{DbSfavored} = 0.91$, respectively; Figure 2.4a), and these results were confirmed at the individual level where DbS was strongly favored in 6 out of 6 participants in both the perceptual and value conditions (Figure 2.4—figure supplement 1). Once again, we found that inclusion of choice history biases in this experiment did not significantly affect our results both at the population and individual levels. Population probability that DbS was deemed best in the perceptual ($p_{DbSfavored} = 0.999$) and value ($p_{DbSfavored} = 0.90$) conditions (Figure

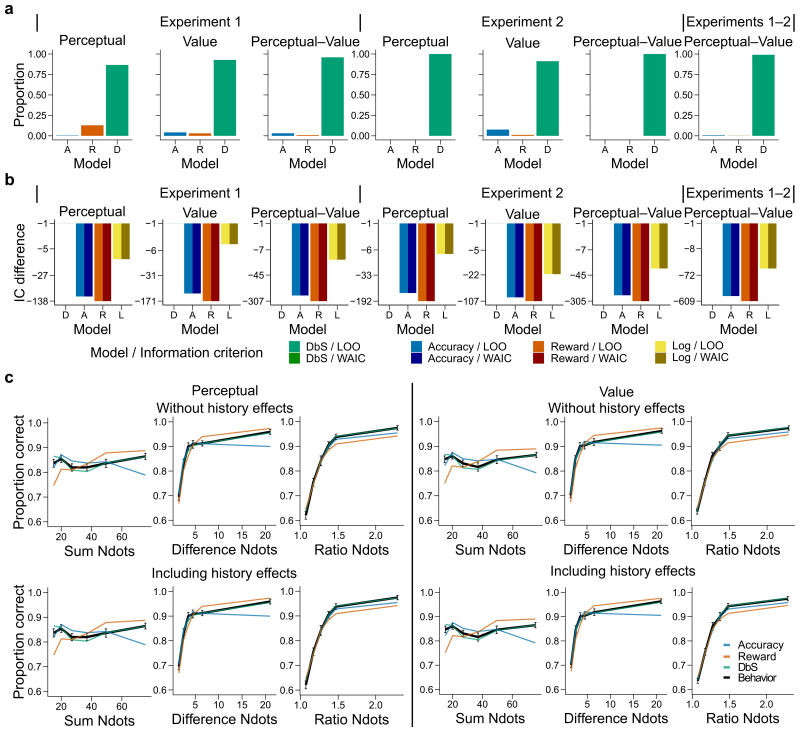


Figure 2.4: **Behavioral results. (a)** Bars represent proportion of times an encoding rule (Accuracy [A, blue], Reward [R, red], DbS [D, green]) was selected by the Bayesian latent-mixture model based on the posterior estimates across participants. Each panel shows the data grouped for each and across experiments and experimental conditions (see titles on top of each panel). The results show that DbS was clearly the favored encoding rule. The latent vector π posterior estimates are presented in Figure 2.4 - figure supplement 4. **(b)** Difference in LOO and WAIC between the best model (DbS (D) in all cases) and the competing models: Accuracy (A), Reward (R) and Logarithmic (L) models. Each panel shows the data grouped for each and across experimental conditions and experiments (see titles on top of each panel). **(c)** Behavioral data (black, error bars represent SEM across participants) and model predictions based on fits to the empirical data. Data and model predictions are presented for both the perceptual (left panels) or value (right panels) conditions, and excluding (top panels) or including (bottom panels) choice history effects. Performance of data model predictions is presented as function of the sum of the number of dots in both clouds (left), the absolute difference between the number of dots in both clouds (middle) and the ratio of the number of dots in the most numerous cloud over the less numerous cloud (right). Results reveal a remarkable overlap of the behavioral data and predictions by DbS, thus confirming the quantitative results presented in panels a and b.

2.4—figure supplement 1), and at the individual level DbS was strongly favored in 6 out of 6 participants in the perceptual condition and 5 of 6 in the value condition (Figure 2.4—figure supplement 1). Thus, Experiments 1 and 2 strongly suggest that our results are not driven by specific instructions or characteristics of the behavioral task.

As a further robustness check, for each participant we grouped the data in different ways across experiments (Experiments 1 and 2) and experimental conditions (perceptual or value) and investigated which sampling model was favored. We found that irrespective of how the data was grouped, DbS was the model that was clearly deemed best at the population (Figure 2.4) and individual level (Figure 2.4—figure supplement 3). Additionally, we investigated whether these quantitative results specifically depended on our choice of using a latent-mixture model. Therefore, we also fitted each model independently and compared the quality of the model fits based on out-of-sample cross-validation metrics (Methods). Once again, we found that the DbS model was favored independently of experiment and conditions (Figure 2.4).

One possible reason why the two experimental conditions did not lead to differences could be that, after doing one condition for two days, the participants did not adapt as easily to the new incentive rule. However, note that as the participants did not know of the second condition before carrying it out, they could not adopt a compromise between the two behavioral objectives. Nevertheless, we fitted the latent-mixture model only to the first condition that was carried out by each participant. We found once again that DbS was the best model explaining the data, irrespective of condition and experimental paradigm (Figure 2.4—figure supplement 7). Therefore, the fact that DbS is favored in the results is not an artifact of carrying out two different conditions in the same participants.

We also investigated whether the DbS model makes more accurate predictions than the widely used logarithmic model of numerosity discrimination tasks [49]. We found that DbS still made better out-of-sample predictions than the log-model (Figure 2.4b, Figure 2.5f,g). Moreover, these results continued to hold after taking into account possible choice history biases (Figure 2.4—figure supplement 4). In addition to these quantitative results, qualitatively we also found that behavior closely matched the predictions of the DbS model remarkably well (Figure 2.4c), based on virtually only one free parameter, namely, the number of samples (resources) n. Together, these results provide

compelling evidence that DbS is the most likely resource-constrained sampling strategy used by participants in numerosity discrimination tasks.

Recent studies have also investigated behavior in tasks where perceptual and preferential decisions have been investigated in paradigms with identical visual stimuli [101–103]. In these tasks, investigators have reported differences in behavior, in particular in the reaction times of the responses, possibly reflecting differences in behavioral strategies between perceptual and value-based decisions. Therefore, we investigated whether this was the case also in our data. We found that reaction times did not differ between experimental conditions for any of the different performance assessments considered here (Figure 2.4—figure supplement 5). This further supports the idea that participants were in fact using the same sampling mechanism irrespective of behavioral goals.

Here it is important to emphasize that all sampling models and the logarithmic model of numerosity have the same degrees of freedom (performance is determined by n in the sampling models and Weber's fraction σ in the log model, Methods). Therefore, qualitative and quantitative differences favoring the DbS model cannot be explained by differences in model complexity. It could also be argued that normal approximation of the binomial distributions in the sampling decision models only holds for large enough n. However, we find evidence that the large-n optimal solutions are also nearly optimal for low n values (Appendix 2.7). Estimates of n in our data are in general $n \approx 21$ (Table 1) and we find that the large-n rule is nearly optimal already for $n = 15$ (Appendix 2.7). Therefore the asymptotic approximations should not greatly affect the conclusions of our work.

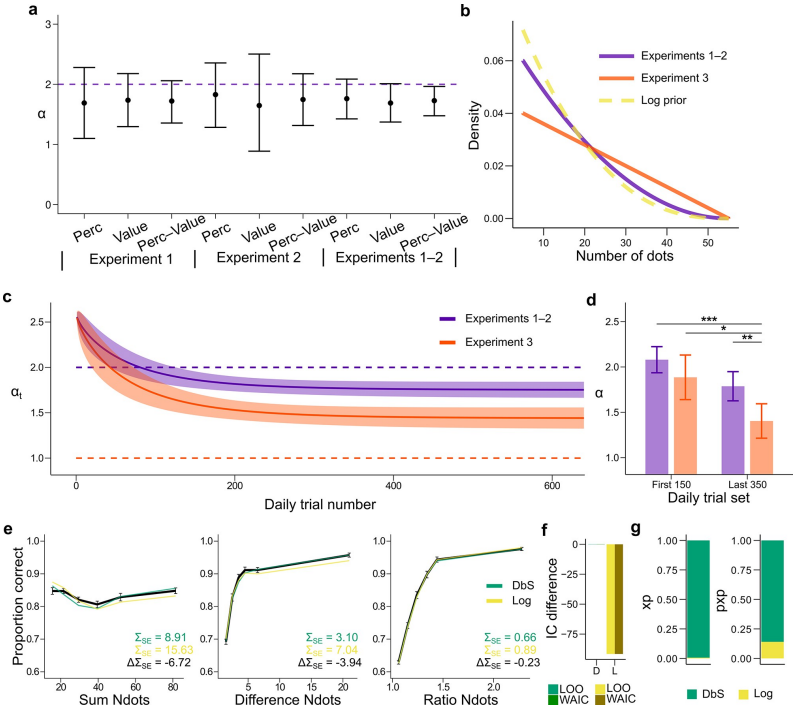


Figure 2.5: **Prior adaptation analyses. (a)** Estimation of the shape parameter α for the DbS model by grouping the data for each and across experimental conditions and experiments. Error bars represent the 95% highest density interval of the posterior estimate of α at the population level. The dashed line shows the theoretical value of α. **(b)** Theoretical prior distribution $f(v)$ in Experiments 1 and 2 ($\alpha = 2$, purple) and 3 ($\alpha = 1$, orange). The dashed line represents the value of α of our prior parametrization that approximates the DbS and log discriminability models. **(c)** Posterior estimation of α_t (Equation 18) as a function of the number of trials t in each daily session for Experiments 1 and 2 (purple) and Experiment 3 (orange). The results reveal that, as expected, α_t reaches a lower asymptotic value δ. Error bars represent ± SD of 3000 simulated α_t values drawn from the posterior estimates of the HBM (see Materials and methods). **(d)** Model fit to the first 150 and last 350 trials of each daily session. The α parameter was allowed to vary between the first and last sets of daily trials and between Experiments 1–2 and Experiment 3. In Experiment 3, α is lower in the last set of trials compared to the first set of trials (PMCMC=0.013). In addition, α for the last trials is lower for Experiment 3 than for Experiments 1–2 (PMCMC=0.006). This confirms that the results presented in panel c are not artifacts of the adaptation parametrization assumed for α. Error bars represent ± SD of the posterior chains of the corresponding parameter. (*P<0.05, **P<0.01, and ***P<0.001).

Figure 2.5:
(continued) **(e)** Behavioral data (black) and model fit predictions of the DbS (green) and Log (yellow) models. Performance of each model as a function of the sum of the number of dots in both clouds (left), the absolute difference between the number of dots in both clouds (middle) and the ratio of the number of dots in the most numerous cloud over the less numerous cloud (right). Error bars represent SEM **(f)** Difference in LOO and WAIC between the best fitting DbS (D) and logarithmic encoding (Log) model. **(g)** Population exceedance probabilities (xp, left) and protected exceedance probabilities (pxp, right) for DbS (green) vs Log (yellow) of a Bayesian model selection analysis [100]: $xp_{DbS} = 0.99$, $pxp_{DbS} = 0.87$. These results provide a clear indication that the adaptive DbS explains the data better than the Log model.

Dynamics of adaptation

Up to now, fits and comparison across models have been done under the assumption that the participants learned the prior distribution $f(v)$ imposed in our task. If participants are employing DbS, it is important to understand the dynamical nature of adaptation in our task. Note that the shape of the prior distribution is determined by the parameter α (Figure 2.5b, Equation 10 in Methods). First, we made sure based on model recovery analyses that the DbS model could jointly and accurately recover both the shape parameter α and the resource parameter n based on synthetic data (Figure 2.3—figure supplement 2). Then we fitted this model to the empirical data and found that the recovered value of the shape parameter α closely followed the value of the empirical prior with a slight underestimation (Figure 2.5a). Next, we investigated the dynamics of prior adaptation. To this end, we ran a new experiment (Experiment 3, n = 7 new participants) in which we set the shape parameter of the prior to a lower value compared to Experiments 1–2 (Figure 2.5b, Methods). We investigated the change of α over time by allowing this parameter to change with trial experience (Equation 18, Methods) and compared the evolution of α for Experiments 1 and 2 (empirical $\alpha = 2$) with Experiment 3 (empirical $\alpha = 1$, Figure 2.5b). If participants show prior adaptation in our numerosity discrimination task, we hypothesized that the asymptotic value of α should be higher for Experiments 1–2 than for Experiment 3. First, we found that for Experiments 1–2, the value of α quickly reached an asymptotic value close to the target value (Figure 2.5c). On the other hand, for Experiment 3 the value of α continued to decrease during the experimental session, but slowly approaching its target value. This seemingly slower adaptation to the shape of the prior in Experiment 3 might be explained by the following observation. The prior parametrized with $\alpha = 1$ in Experiment 3 is further away from an agent hypothesized to have a natural numerosity discrimination based on a log scale ($\alpha = 2.58$,

Experiment	Condition	History effects	Model		
			$n_{Accuracy}$	n_{Reward}	n_{DbS}
1	V	not included	15.24 ± 3.09	17.54 ± 3.98	24.40 ± 5.16
2	V	not included	22.48 ± 2.43	27.58 ± 3.81	35.40 ± 3.44
1	P	not included	15.19 ± 3.99	17.84 ± 4.85	24.64 ± 6.59
2	P	not included	20.99 ± 1.59	24.22 ± 1.93	33.54 ± 2.45
1	P/V	not included	15.33 ± 3.41	17.25 ± 4.45	24.15 ± 5.75
2	P/V	not included	21.30 ± 0.96	25.27 ± 1.99	33.90 ± 1.51
1/2	V	not included	18.56 ± 2.04	22.05 ± 2.73	29.52 ± 3.25
1/2	P	not included	17.91 ± 2.09	20.66 ± 2.59	28.62 ± 3.51
1/2	P/V	not included	17.93 ± 1.87	21.03 ± 2.46	28.58 ± 3.04
1	V	included	15.50 ± 3.13	17.50 ± 3.91	24.68 ± 5.08
2	V	included	22.92 ± 2.37	28.07 ± 3.73	36.18 ± 2.91
1	P	included	15.41 ± 3.81	17.96 ± 4.88	24.70 ± 6.62
2	P	included	21.57 ± 1.71	24.88 ± 2.17	34.37 ± 2.93
1	P/V	included	15.16 ± 3.55	17.43 ± 4.39	24.30 ± 5.94
2	P/V	included	21.80 ± 0.92	25.81 ± 1.86	34.60 ± 1.40
1/2	V	included	18.86 ± 2.07	22.48 ± 2.75	29.85 ± 3.17
1/2	P	included	18.15 ± 2.17	21.11 ± 2.72	29.01 ± 3.47
1/2	P/V	included	18.22 ± 1.93	21.34 ± 2.50	29.12 ± 3.12

Table 2.1: **Resource parameter n fits.** Fits of the resource parameter for the Accuracy, Reward and Decision by Sampling (DbS) models combining data across experiments and conditions (Perceptual (P) or Value (V)) either including or ignoring history effects. The values represent the mean ± SD of the population mean of the number of resources. To fit the same behavior data, the Reward and in particular the DbS models require a higher number of resources than the Accuracy model, which is coherent with the fact that the Accuracy model allocates its resources to maximize accuracy, therefore reducing the number of resources needed to reach a given accuracy.

Methods), which is closer in value to the shape of the prior in Experiments 1 and 2 ($\alpha = 2$). Irrespective of these considerations, the key result to confirm our adaptation hypothesis is that the asymptotic value of α is lower for Experiment 3 compared to Experiments 1 and 2 ($P_{MCMC} = 0.006$).

In order to make sure that this result was not an artifact of the parametric form of adaptation assumed here (Equation 18, Methods), we fitted the DbS model to trials at the beginning and end of each experimental session allowing α to be a free but fixed parameter in each set of trials. The results of these new analyses are virtually identical to the results obtained with the parametric form, in which α is smaller at the end of Experiment 3

sessions relative to beginning of Experiments 1 and 2 ($P_{MCMC} = 0.0003$), beginning of Experiments 3 ($P_{MCMC} = 0.013$) and end of Experiments 1 and 2 ($P_{MCMC} = 0.006$, Figure 2.5d). In this model, we did not allow n to freely change for each condition, and therefore a concern might be that the results might be an artifact of changes in n, which could for example change with the engagement of the participants across the session. Given that we already demonstrated that both parameters n and α are identifiable, we fitted the same model as in Figure 2.5d, however this time we allowed n to be free parameter alongside α. We found that the results obtained in Figure 2.5d remained virtually unchanged (Figure 2.5—figure supplement 3), in addition to the result that the resource parameter n remained virtually identical across the session (Figure 2.5—figure supplement 3).

We further investigated evidence for adaptation using an alternative quantitative approach. First, we performed out-of-sample model comparisons based on the following models: (i) the adaptive-α model, (ii) free-α model with α free but non-adapting over time, and (iii) fixed-α model with $\alpha = 2$. The results of the out-of-sample predictions revealed that the best model was the free-α model, followed closely by the adaptive-α model ($\Delta LOO = 1.8$) and then by fixed-α model ($\Delta LOO = 32.6$). However, we did not interpret the apparent small difference between the adaptive-α and the free-α models as evidence for lack of adaptation, given that the more complex adaptive-α model will be strongly penalized after adaptation is stable. That is, if adaptation is occurring, then the adaptive-α only provides a better fit for the trials corresponding to the adaptation period. After adaptation, the adaptive-α should provide a similar fit than the free-α model, however with a larger complexity that will be penalized by model comparison metrics. Therefore, to investigate the presence of adaptation, we took a closer quantitative look at the evolution of the fits across trial experience. We computed the average trial-wise predicted Log-Likelihood (by sampling from the hierarchical Bayesian model) and compared the differences of this metric between the competing models and the adaptive model. We hypothesized that if adaptation is taking place, the adaptive-α model would have an advantage relative to the free-α model at the beginning of the session, with these differences vanishing toward the end. On the other hand, the fixed-α should roughly match the adaptive-α model at the beginning and then become worse over time, but these differences should stabilize after the end of the adaptation period. The results of these analyses support our hypotheses (Figure 2.5—figure supplement 2), thus providing further evidence of adaptation, highlighting the fact that the DbS model can parsimoniously capture adaptation to contextual changes in a continuous

and dynamical manner. Furthermore, we found that the DbS model again provides more accurate qualitative and quantitative out-of-sample predictions than the log model (Figure 2.5e,f).

2.4. Discussion

The brain is a metabolically expensive inference machine [7, 104, 105]. Therefore, it has been suggested that evolutionary pressure has driven it to make productive use of its limited resources by exploiting statistical regularities [24, 67, 70]. Here, we incorporate this important —often ignored —aspect in models of behavior by introducing a general framework of decision-making under the constraints that the system: (i) encodes information based on binary codes, (ii) has limited number of samples available to encode information, and (iii) considers the costs of contextual adaptation.

Under the assumption that the organism has fully adapted to the statistics in a given context, we show that the encoding rule that maximizes mutual information is the same rule that maximizes decision accuracy in two-alternative decision tasks. However, note that there is nothing privileged about maximizing mutual information, as it does not mean that the goals of the organism are necessarily achieved [92, 106]. In fact, we show that if the goal of the organism is instead to maximize the expected value of the chosen options, the system should not rely on maximizing information transmission and must give up a small fraction of precision in information coding. Here, we derived analytical solution for each of these optimization objective criteria, emphasizing that these analytical solutions were derived for the large-n limiting case. However, we have provided evidence that these solutions continue to be more efficient relative to DbS for small values of n, and more importantly, they remain nearly optimal even at relatively low values of n, in the range of values that might be relevant to explain human experimental data (Appendix 2.7).

Another key implication of our results is that we provide an alternative explanation to the usual conception of noise as the main cause of behavioral performance degradation, where noise is usually artificially added to models of decision behavior to generate the desired variability [107, 108]. On the contrary, our work makes it formally explicit why a system that evolved to encode information based on binary codes must be necessarily noisy, also revealing how the system could take advantage of its unavoidable noisy properties [109] to optimize decision behavior [110]. Here, it is important to

highlight that this conclusion is drawn from a purely homogeneous neural circuit, in other words, a circuit in which all neurons have the same properties (in our case, the same activation thresholds). This is not what is typically observed, as neural circuits are typically very heterogeneous. However, in the neural circuit that we consider here, it could mean that the firing thresholds can vary across neurons [111], which could be used by the system to optimize the required variability of binary neural codes. Interestingly, it has been shown in recent work that stochastic discrete events also serve to optimize information transmission in neural population coding [94, 112, 113]. Crucially, in our work we provide a direct link of the necessity of noise for systems that aim at optimizing decision behavior under our encoding and limited-capacity assumptions, which can be seen as algorithmic specifications of the more realistic population coding specifications mentioned above [94]. We argue that our results may provide a formal intuition for the apparent necessity of noise for improving training and learning performance in artificial neural networks [114, 115], and we speculate that an implementation of 'the right' noise distribution for a given environmental statistic could be seen as a potential mechanism to improve performance in capacity-limited agents generally speaking [116]. We acknowledge that based on the results of our work, we cannot confirm whether this is the case for higher order neural circuits, however, we leave it as an interesting theoretical formulation, which could be addressed in future work.

Interestingly, our results could provide an alternative explanation of the recent controversial finding that dynamics of a large proportion of LIP neurons likely reflect binary (discrete) coding states to guide decision behavior [117, 118]. Based on this potential link between their work and ours, our theoretical framework generates testable predictions that could be investigated in future neurophysiological work. For instance, noise distribution in neural circuits should dynamically adapt according to the prior distribution of inputs and goals of the organism. Consequently, the rate of 'step-like' coding in single neurons should also be dynamically adjusted (perhaps optimally) to statistical regularities and behavioral goals.

Our results are closely related to Decision by Sampling (DbS), which is an influential account of decision behavior derived from principles of retrieval and memory comparison by taking into account the regularities of the environment, and also encodes information based on binary codes [81]. We show that DbS represents a special case of our more general efficient sampling framework, that uses a rule that is similar to (though not exactly like) the optimal encoding rule that assumes full (or costless) adaptation to the prior statistics

of the environment. In particular, we show that DbS might well be the most efficient sampling algorithm, given that a reduction in the full representation of the prior distribution might not come at a great loss in performance. Interestingly, our experimental results (discussed in more detail below) also provide support for the hypothesis that numerosity perception is efficient in this particular way. Crucially, DbS automatically adjusts the encoding in response to changes in the frequency distribution from which exemplars are drawn in approximately the right way, while providing a simple answer to the question of how such adaptation of the encoding rule to a changing frequency distribution occurs, at a relatively low cost.

On a related line of work, Bhui and Gershman [119] develop a similar, but different specification of DbS, in which they also consider only a finite number of samples that can be drawn from the prior distribution to generate a percept, and ask what kind of algorithm would be required to improve coding efficiency. However, their implementation differs from ours in various important ways (see Appendix 2.8 for a detailed discussion). One of the main distinctions is that they consider the case in which only a finite number of samples can be drawn from the prior and show that a variant of DbS with kernel-smoothing is superior to its standard version. However, a key difference to our implementation is that they allow the kernel-smoothed quantity (computed by comparing the input v with a sample $\hat{v}$ from the prior distribution) to vary continuously between 0 and 1, rather than having to be either 0 or 1 as in our implementation (Figure 2.1). Thus, they show that coding efficiency can be improved by allowing a more flexible implementation of the coding scheme for the case when the agent is allowed to draw few samples from the prior distribution (Appendix 2.8). On the other hand, we restrict our framework to a coding scheme that is only allowed to encode information based on zeros or ones, where we show that coding efficiency can be improved relative to DbS only under a more complete knowledge of the prior distribution, where the optimal solutions can be formally derived in the large-n limit. Nevertheless, we have shown that even under the operation of few sampling units, the optimal rules will be still superior to the standard DbS (if the agent has fully adapted to the statistics of the environment in a given context), even when a few number of processing units are available to generate decision relevant percepts.

We tested these resource-limited coding frameworks in non-symbolic numerosity discrimination, a fundamental cognitive function for behavior in humans and other animals, which may have emerged during evolution to support fitness maximization [48]. Here, we find that the way in which the precision of

numerosity discrimination varies with the size of the numbers being compared is consistent with the hypothesis that the internal representations on the basis of which comparisons are made are sample-based. In particular, we find that the encoding rule varies depending on the frequency distribution of values encountered in a given environment, and that this adaptation occurs fairly quickly once the frequency distribution changes.

This adaptive character of the encoding rule differs, for example, from the common hypothesis of a logarithmic encoding rule (independent of context), which we show fits our data less well. Nonetheless, we can reject the hypothesis of full optimality of the encoding rule for each distribution of values used in our experiments, even after participants have had extensive experience with a given distribution. Thus, a possible explanation of why DbS is the favored model in our numerosity task is that accuracy and reward maximization requires optimal adaptation of the noise distribution based on our imposed prior, requiring complex neuroplastic changes to be implemented, which are in turn metabolically costly [120]. Relying on samples from memory might be less metabolically costly as these systems are plastic in short time scales, and therefore a relatively simpler heuristic to implement allowing more efficient adaptation. Here, it is important to emphasize, as it has been discussed in the past [121, 122], that for decision-making systems beyond the perceptual domain, the identity of the samples is unclear. We hypothesize, that information samples derive from the interaction of memory on current sensory evidence depending on the retrieval of relevant samples to make predictions about the outcome of each option for a given behavioral goal (therefore also depending on the encoding rule that optimizes a given behavioral goal).

Interestingly, it was recently shown that in a reward learning task, a model that estimates values based on memory samples from recent past experiences can explain the data better than canonical incremental learning models [123]. Based on their and our findings, we conclude that sampling from memory is an efficient mechanism for guiding choice behavior, as it allows quick learning and generalization of environmental contexts based on recent experience without significantly sacrificing behavioral performance. However, it should be noted that relying on such mechanisms alone might be suboptimal from a performance- and goal-based point of view, where neural calibration of optimal strategies may require extensive experience, possibly via direct interactions between sensory, memory and reward systems [124, 125].

Taken together, our findings emphasize the need of studying optimal models, which serve as anchors to understand the brain's computational goals without ignoring the fact that biological systems are limited in their capacity to process information. We addressed this by proposing a computational problem, elaborating an algorithmic solution, and proposing a minimalistic implementational architecture that solves the resource-constrained problem. This is essential, as it helps to establish frameworks that allow comparing behavior not only across different tasks and goals, but also across different levels of description, for instance, from single cell operation to observed behavior [126]. We argue that this approach is fundamental to provide benchmarks for human performance that can lead to the discovery of alternative heuristics [127, 128] that could appear to be in principle suboptimal, but that might be in turn the optimal strategy to implement if one considers cognitive limitations and costs of optimal adaptation. We conclude that the understanding of brain function and behavior under a principled research agenda, which takes into account decision mechanisms that are biologically feasible, will be essential to accelerate the elucidation of the mechanisms underlying human cognition.

2.5. Methods

Participants

The study tested young healthy volunteers with normal or corrected-to-normal vision (total n = 20, age 19–36 years, nine females: n = 7 in Experiment 1, two females; n = 6 new participants in Experiment 2, three females; n = 7 new participants in Experiment 3, four females). Participants were randomly assigned to each experiment and no participant was excluded from the analyses. Participants were instructed about all aspects of the experiment and gave written informed consent. None of the participants suffered from any neurological or psychological disorder or took medication that interfered with participation in our study. Participants received monetary compensation for their participation in the experiment partially related to behavioral performance (see below). The experiments conformed to the Declaration of Helsinki and the experimental protocol was approved by the Ethics Committee of the Canton of Zurich (BASEC: 2018–00659).

Experiment 1

Participants (n = 7) carried out a numerosity discrimination task for four consecutive days for approximately one hour per day. Each daily session consisted of a training run followed by 8 runs of 75 trials each. Thus, each participant completed ∼2400 trials across the four days of experiment.

After a fixation period (1–1.5 s jittered), two clouds of dots (left and right) were presented on the screen for 200 ms. Participants were asked to indicate the side of the screen where they perceived more dots. Their response was kept on the screen for 1 s followed by feedback consisting of the symbolic number of dots in each cloud as well as the monetary gains and opportunity losses of the trial depending on the experimental condition. In the value condition, participants were explicitly informed that each dot in a cloud of dots corresponded to 1 Swiss Franc (CHF). Participants were informed that they would receive the amount in CHF corresponding to the total number of dots on the chosen side. At the end of the experiment a random trial was selected and they received the corresponding amount. In the accuracy condition, participants were explicitly informed that they could receive a fixed reward (15 Swiss Francs (CHF)) for each correct trial. This fixed amount was selected such that it approximately matched the expected reward received in the value condition (as tested in pilot experiments). At the end of the experiment, a random trial was selected and they would receive this fixed amount if they chose the cloud with more dots (i.e., the correct side). Each condition lasted for two consecutive days with the starting condition randomized across participants. Only after completing all four experiment days, participants were compensated for their time with 20 CHF per hour, in addition to the money obtained based on their decisions on each experimental day.

Experiment 2

Participants (n = 6) carried out a numerosity discrimination task in which each of four daily sessions consisted of 16 runs of 40 trials each, thus each participant completed ∼ 2560 trials. A key difference with respect to Experiment 1 is that participants had to accumulate points based on their decisions and had to reach a predetermined threshold on each run. The rules of point accumulation depended on the experimental condition. In the perceptual condition, a fixed amount of points was awarded if the participants chose the cloud with more dots. In this condition, participants were instructed to accumulate a number of points and reach a threshold given a limited number of trials. Based on the results obtained in Experiment 1, the threshold corre-

sponded to 85% of correct trials in a given run, however the participants were unaware of this. If the participants reached this threshold, they were eligible for a fixed reward (20 CHF) as described in Experiment 1. In the value condition, the number of points received was equal to the number of dots in the cloud, however, contrary to Experiment 1, points were only awarded if the participant chose the cloud with the most dots. Participants had to reach a threshold that was matched in the expected collection of points of the perceptual condition. As in Experiment 1, each condition lasted for two consecutive days with the starting condition randomized across participants. Only after completing all the four days of the experiment, participants were compensated for their time with 20 CHF per hour, in addition to the money obtained based on their decisions on each experimental day.

Experiment 3

The design of Experiment 3 was similar to the value condition of Experiment 2 (n = 7 participants) and was carried out over three consecutive days. The key difference between Experiment 3 and Experiments 1–2 was the shape of the prior distribution $f(v)$ that was used to draw the number of dots for each cloud in each trial (see below).

Stimuli statistics and trial selection

For all experiments, we used the following parametric form of the prior distribution

$$f(v) = c(1 - v)^{\alpha} \tag{2.10}$$

initially defined in the interval [0,1] for mathematical tractability in the analytical solution of the encoding rules $\theta(v)$ (see below), with $\alpha > 0$ determining the shape of the distribution, and c is a normalizing constant. For Experiments 1 and 2 the shape parameter was set to $\alpha = 2$, and for Experiment 3 was set to $\alpha = 1$. i.i.d. samples drawn from this distribution were then multiplied by 50, added an offset of 5, and finally were rounded to the closest integer (i.e., the numerosity values in our experiment ranged from $v_{min} = 5$ to $v_{max} = 55$). The pairs of dots on each trial were determined by sampling from a uniform density window in the CDF space (Equation 10 is its corresponding PDF). The pairs of dots in each trial were selected with the conditions that, first, their distance in the CDF space was less than a constant (0.25, 0.28 and 0.23 for Experiments 1, 2 and 3 respectively),

and second, the number of dots in both clouds was different. Figure 2.3c illustrates the probability that a pair of choice alternatives was selected for a given trial in Experiments 1 and 2.

Power analyses and model recovery

Given that adaptation dynamics in sensory systems often require long-term experience with novel prior distributions, we opted for maximizing the number of trials for a relatively small number of participants per experiment, as it is commonly done for this type of psychophysical experiments [129–131]. Note that based on the power analyses described below, we collected in total ∼45,000 trials across the three Experiments, which is above the average number of trials typically collected in human studies.

In order to maximize statistical power in the differentiation of the competing encoding rules, we generated 10,000 sets of experimental trials for each encoding rule and selected the sets of trials with the highest discrimination power (i.e., largest differences in Log-Likelihood) between the encoding models. In these power analyses, we also investigated what was the minimum number of trials that would allow accurate generative model selection at the individual level. We found that ∼1000 trials per participant in each experimental condition would be sufficient to predict accurately ($P > 0.95$) the true generative model. Based on these analyses, we decided to collect at least 1200 trials per participant and condition (perceptual and value) in each of the three experiments. Model recovery analyses presented in Figure 2.3d illustrate the result of our power analyses (see also Figure 2.3—figure supplement 1).

Apparatus

Eyetracking (EyeLink 1000 Plus) was used to check the participants' fixation during stimulus presentation. When participants blinked or moved their gaze (more than 2° of visual angle) away from the fixation cross during the stimulus presentation, the trial was canceled (only 212 out of 45,600 trials were canceled, that is, <0.5% of the trials). Participants were informed when a trial was canceled and were encouraged not to do so as they would not receive any reward for this trial. A chinrest was used to keep the distance between the participants and the screen constant (55 cm). The task was run using Psychtoolbox Version 3.0.14 on Matlab 2018a. The diameter of the dots varied between 0.42° and 1.45° of visual angle. The center of each cloud was positioned 12.6° of visual angle horizontally from the fixation cross

and had a maximum diameter of 19.6° of visual angle. Following previous numerosity experiments [132, 133], either the average dot size or the total area covered by the dots was maintained constant in both clouds for each trial. The color of each dot (white or black) was randomly selected for each dot. Stimuli sets were different for each participant but identical between the two conditions.

Encoding rules and model fits

The parametrization of the prior $f(v)$ (Equation 10) allows tractable analytical solutions of the encoding rules $\theta_A(v)$, $\theta_R(v)$ and $\theta_D(v)$, that correspond to Accuracy maximization, Reward maximization, and DbS, respectively:

$$\theta_A(v) = sin\left[\frac{\pi}{2}(1-(1-v)^{\alpha+1})\right]^2 \tag{2.11}$$

$$\theta_R(v) = sin\left[\frac{\pi}{2}(1+(v-1)((1-v)^{\alpha})^{2/3})\right]^2 \tag{2.12}$$

$$\theta_D(v) = 1-(1-v)^{\alpha+1} \tag{2.13}$$

Graphical representation of the respective encoding rules is shown in Figure 2.3e for Experiments 1 and 2. Given an encoding rule $theta(v)$, we now define the decision rule. The goal of the decision maker in our task is always to decide which of two input values v_1 and v_2 is larger. Therefore, the agent choses v_1 if and only if the internal readings $k_1 > k_2$. Following the definitions of expected value and variance of binomial variables, and approximating for large n (see Appendix 2.2), the probability of choosing v_1 is given by

$$P_{choosev_1} \approx \Phi\left(\frac{\theta_1-\theta_2}{\sqrt{\frac{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}{n}}}\right) \tag{2.14}$$

where $\Phi()$ is the standard CDF, and θ_1 and θ_2 are the encoding rules for the input values v_1 and v_2, respectively. Thus, the choice structure is the same for all models, only differing in their encoding rule. The three models generate different qualitative performance predictions for a given number of samples n (Figure 2.3f).

Crucially, this probability decision rule (Equation 14) can be parsimoniously extended to include potential side biases independent of the encoding process as follows

$$P_{choosev_1} \approx \Phi\left(\frac{\theta_1 - \theta_2}{\sqrt{\frac{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}{n}}} + \beta_0\right) \tag{2.15}$$

where β_0 is the bias term. This is the base model used in our work. We were also interested in studying whether choice history effects [96, 98] may have influence in our task, thus possibly affecting the conclusions that can be drawn from the base model. Therefore, we extended this model to incorporate the effect of decision learning and choices from the previous trial

$$P_{choosev_1} \approx \Phi\left(\frac{\theta_1 - \theta_2}{\sqrt{\frac{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}{n}}} + \beta_0 + \beta^L a_{t-1} r_{t-1} + \beta^{Ch} a_{t-1}\right) \tag{2.16}$$

where a_{t-1} is the choice made on the previous trial (+1 for left choice and -1 for right choice) and r_{t-1} is the 'outcome learning' on the previous trial (+1 for correct choice and -1 for incorrect choice). β^L and β^{Ch} capture the effect of decision learning and choice in the previous trial, respectively.

Given that the choice structure is the same for all three sampling models considered here, we can naturally address the question of what decision rule the participants favor via a latent-mixture model. We implemented this model based on a hierarchical Bayesian modelling (HBM) approach. The base-rate probabilities for the three different encoding rules at the population level are represented by the vector π, so that π_m is the probability of selecting encoding rule model m. We initialize the model with an uninformative prior given by

$$\pi \sim Dirichlet(1_{m=1}, 1_{m=2}, 1_{m=3}).$$

This base-rate is updated based on the empirical data, where we allow each participant s to draw from each model categorically based on the updated base-rate

$$m_s \sim Categorial(\pi),$$

where the encoding rule θ for model m is given by

$$\theta_{m,s} = \begin{cases} \theta_A, & m = 1 \\ \theta_R, & m = 2 \\ \theta_D, & m = 3 \end{cases}$$

The selected rule was then fed into Equations 15 and 16 to determine the probability of selecting a cloud of dots. The number of samples n was also estimated within the same HBM with population mean μ and standard deviation σ initialized based on uninformative priors with plausible ranges

$$\mu_n \sim Uniform(1, 1000)$$
$$\sigma_n \sim Uniform(0.01, 1000)$$

allowing each participant s to draw from this population prior assuming that n is normally distributed at the population level

$$n_s \sim Normal(\mu_n, \sigma_n)$$

Similarly, the latent variables β in equations Equations 15 and 16 were estimated by setting population mean μ_β and standard deviation σ_β initialized based on uninformative priors

$$\mu_\beta \sim Uniform(-10, 10)$$
$$\sigma_\beta \sim Uniform(0.01, 100)$$

allowing each participant s to draw from this population prior assuming that β is normally distributed at the population level

$$\beta_s \sim Normal(\mu_\beta, \sigma_\beta)$$

In all the results reported in Figure 2.3 and Figure 2.4, the value of the shape parameter of the prior was set to its true value $\alpha = 2$. The estimation of α in Figure 2.5a was investigated with a similar hierarchical approach, allowing each participant to sample from the normal population distribution with uninformative priors over the population mean and standard deviation

$$\mu_\alpha \sim Uniform(0.01, 20)$$
$$\sigma_\alpha \sim Uniform(0.0001, 100)$$

The choice rule of the standard logarithmic model of numerosity discrimination is given by

$$P_{choosev_1} = \Phi\left(\frac{log(v_1) - log(v_2)}{\sigma\sqrt{2}}\right) \tag{2.17}$$

where σ is the internal noise in the logarithmic space. This model was extended to incorporate bias and choice history effects in the same way as implemented in the sampling models. Here, we emphasize that all sampling and log models have the same degrees of freedom, where performance is mainly determined by n in the sampling models and Weber's fraction σ in the log model, and biases are determined by parameters β. For all above-mentioned models, the trial-by-trial likelihood of the observed choice (i.e., the data) given probability of a decision was based on a Bernoulli process

$$y_{t,s} \sim Bernoulli(P_{choosev_1})$$

where $y_{t,s} \in \{0, 1\}$ is the decision of each participant s in each trial t. In order to allow for prior adaptation, the model fits presented in Figure 2.3 and Figure 2.4 were fit starting after a fourth of the daily trials (corresponding to 150 trials for Experiment 1 and 160 trials for Experiment 2) to allow for prior adaptation and fixing the shape parameter to its true generative value $\alpha = 2$.

The dynamics of adaptation (Figure 2.5) were studied by allowing the shape parameter α to evolve through trial experience using all trials collected on each experiment day. This was studied using the following function

$$\alpha_t = \delta + \eta e^{-t/\tau} \tag{2.18}$$

where δ represents a possible target adaptation value of α, t is the trial number, and η, τ determine the shape of the adaptation. Therefore, the encoding rule of the DbS model also changed trial-to-trial

$$\theta_D^t(v) = 1 - (1 - v)^{\alpha_t + 1} \tag{2.19}$$

Adaptation was tested based on the hypothesis that participants initially use a logarithmic discrimination rule (Equation 17) (this strategy also allowed improving identification of the adaptation dynamics). Therefore, Equation 18 was parametrized such that the initial value of the shape parameter ($\alpha_{t=0}$) guaranteed that discriminability between the DbS and the logarithmic rule was as close as possible. This was achieved by finding the value of α in the DbS encoding rule (θ_D) that minimizes the following expression

$$\sum_{t=1}^{T} \left[\left(\frac{\theta_D(v_{1,t}) - \theta_D(v_{2,t})}{\sqrt{\theta_D(v_{1,t})(1 - \theta_D(v_{1,t})) + \theta_D(v_{2,t})(1 - \theta_D(v_{2,t}))}} \right) - (log(v_{1,t}) - log(v_{2,t})) \right]^2 \tag{2.20}$$

where $v_{1,t}$ and $v_{2,t}$ are the numerosity inputs for each trial t. This expression was minimized based on all trials generated in Experiments 1–3 (note that minimizing this expression does not require knowledge of the sensitivity levels σ and n for the log and DbS models, respectively). We found that the shape parameter value that minimizes Equation 20 is $\alpha = 2.58$. Based on our prior $f(v)$ parametrization (Equation 10), this suggests that the initial prior is more skewed than the priors used in Experiments 1–3 (Figure 2.5b). This is an expected result given that log-normal priors, typically assumed in numerosity tasks, are also highly skewed. We fitted the δ parameter independently for Experiments 1–2 and Experiment 3 but kept the τ parameter shared across all experiments. If adaptation is taking place, we hypothesized that the asymptotic value δ of the shape parameter α should be larger for Experiments 1–2 compared to Experiment 3.

Posterior inference of the parameters in all the hierarchical models described above was performed via the Gibbs sampler using the Markov Chain Monte

Carlo (MCMC) technique implemented in JAGS. For each model, a total of 50,000 samples were drawn from an initial burn-in step and subsequently a total of new 50,000 samples were drawn for each of three chains (samples for each chain were generated based on a different random number generator engine, and each with a different seed). We applied a thinning of 50 to this final sample, thus resulting in a final set of 1000 samples for each chain (for a total of 3000 pooling all three chains). We conducted Gelman–Rubin tests for each parameter to confirm convergence of the chains. All latent variables in our Bayesian models had $\hat{R} < 1.05$, which suggests that all three chains converged to a target posterior distribution. We checked via visual inspection that the posterior population level distributions of the final MCMC chains converged to our assumed parametrizations. When evaluating different models, we are interested in the model's predictive accuracy for unobserved data, thus it is important to choose a metric for model comparison that considers this predictive aspect. Therefore, in order to perform model comparison, we used a method for approximating leave-one-out cross-validation (LOO) that uses samples from the full posterior [134]. These analyses were repeated using an alternative Bayesian metric: the WAIC [134].

Acknowledgments:

This work was supported by an ERC starting grant (ENTRAINER) to R.P. and by a grant of the U.S. National Science Foundation to M.W. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 758604).

Author contributions:

Joseph A. Heng, Conceptualization, Software, Formal analysis, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Michael Woodford, Conceptualization, Formal analysis, Writing - original draft, Writing - review and editing; Rafael Polania, Conceptualization, Resources, Formal analysis, Supervision, Funding acquisition, Investigation, Writing - original draft, Project administration, Writing - review and editing.

Data and materials availability:

Data and essential code that support the findings of this study have been made available at the Open Science Framework (https://osf.io/xgfu9/).

Supplementary Figures

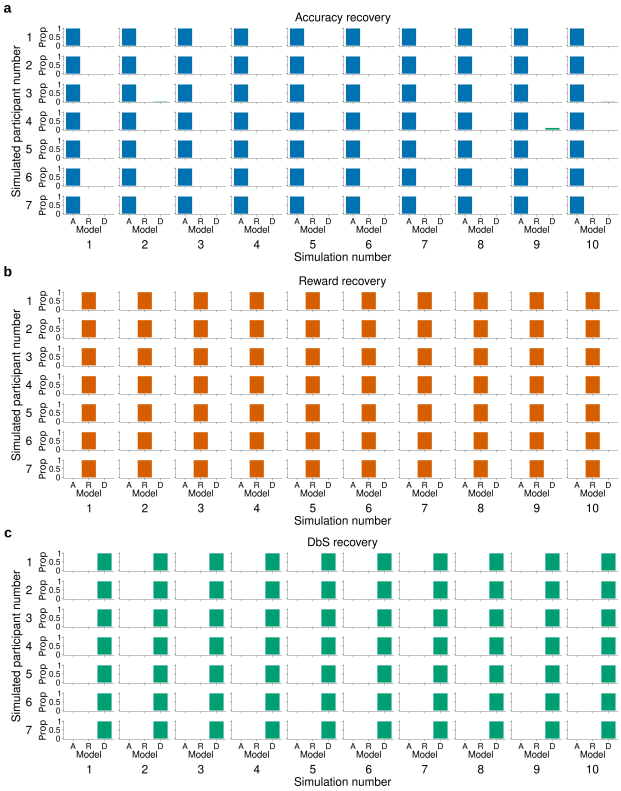


Figure 2.3 - figure supplement 1. Model recovery for α fixed. The latent-mixture model was fitted to synthetic data obtained by simulating 10 times each encoding rule on the trials from participants of Experiment 1. This also means that we used the same number of trials per condition that each participant experienced in our experiments. Each histogram shows the proportion (Prop) of the recovered encoding rule for synthetic data from **(a)** the accuracy maximizing encoding rule θ_A, **(b)** the reward maximizing encoding rule θ_R, and **(c)** decision by sampling θ_D. The latent mixture model can accurately recover the underlying encoding rule. In this model the α parameter was set to 2.

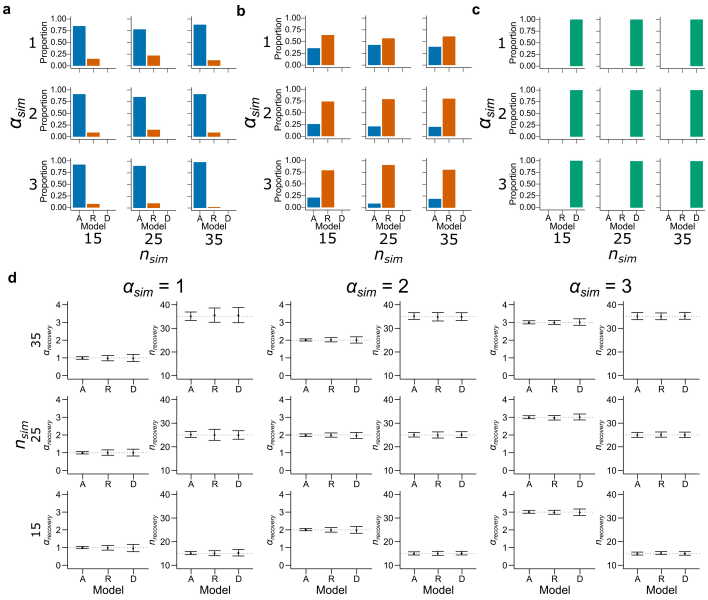


Figure 2.3 - figure supplement 2. Model recovery with both α and n as free parameters. Synthetic data preserving the trial set statistics and number of trials per participant used in Experiment 1 was generated 100 times for each encoding rule with various values of α and n. A model for each encoding rule was fitted to the data using maximum likelihood estimators with α and n as free parameters. The histograms represent the proportion of best fitting models for data generated by **(a)** Accuracy, **(b)** Reward and **(c)** DbS models. Results are shown for different simulated values of α (top: $\alpha_{sim} = 1$, middle: $\alpha_{sim} = 2$ and bottom: $\alpha_{sim} = 3$) and n (left: $n_{sim} = 15$, middle: $n_{sim} = 25$ and right: $n_{sim} = 35$). While DbS is always well recovered, the Accuracy and Reward models tend to be confounded with each other. **(d)** This same synthetic data were fitted with its generating model with α and n as free parameters using maximum likelihood estimators. Results are shown for different simulated values of α (first and second columns: $\alpha_{sim} = 1$, third and fourth columns: $\alpha_{sim} = 2$ and fifth and sixth columns: $\alpha_{sim} = 3$) and n (top: $n_{sim} = 35$, middle: $n_{sim} = 25$ and bottom: $n_{sim} = 15$). Error bars represent ± SD of the recovered parameter across simulations. The parameters are well recovered by the respective generating model.

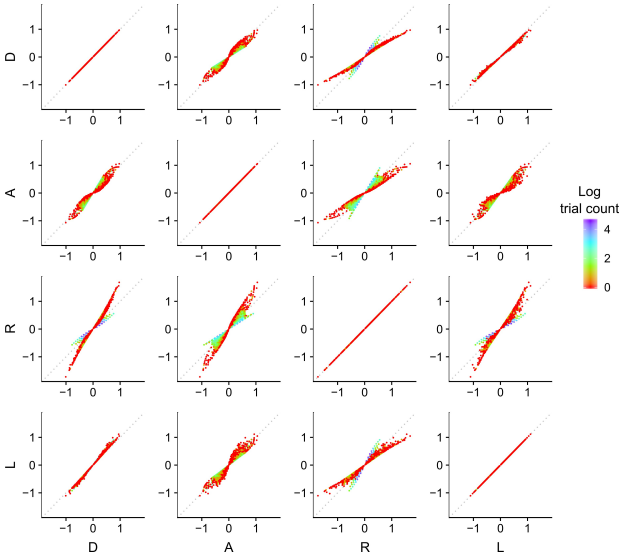


Figure 2.3 - figure supplement 3. Discriminability differences between the different encoding rules.

This figure illustrates the discriminability differences between the different encoding rules considered in this study. Each dot represents the discriminability value for a pair of numerosity values v_1 and v_2 presented on a given trial to the participants in Experiment 1. For the sampling models, the discriminability rule is defined as

$$\frac{\theta(v_1) - \theta(v_2)}{\sqrt{\theta(v_1)(1-\theta(v_1)) + \theta(v_2)(1-\theta(v_2))}},$$

where θ corresponds to the respective Accuracy maximizing (A), Reward maximizing (R) or Decision by Sampling (D) encoding rules. For the logarithmic model (L) the discriminability rule is defined as

$$log(v1) - log(v2).$$

The color of each dot represents the log of the number of occurrences for the pairs of input values v_1 and v_2. Note that the encoding values of the presented numerosities are different depending on the encoding rule, which makes it possible to identify the participants' encoding strategy. Also note that for our imposed prior distribution, the DbS encoding rule is similar to the logarithmic

rule, which explains the smaller difference in the quantitative predictions between these two models. Nevertheless, DbS was always the model that provided the best quantitative and qualitative predictions irrespective of incentivized goals.

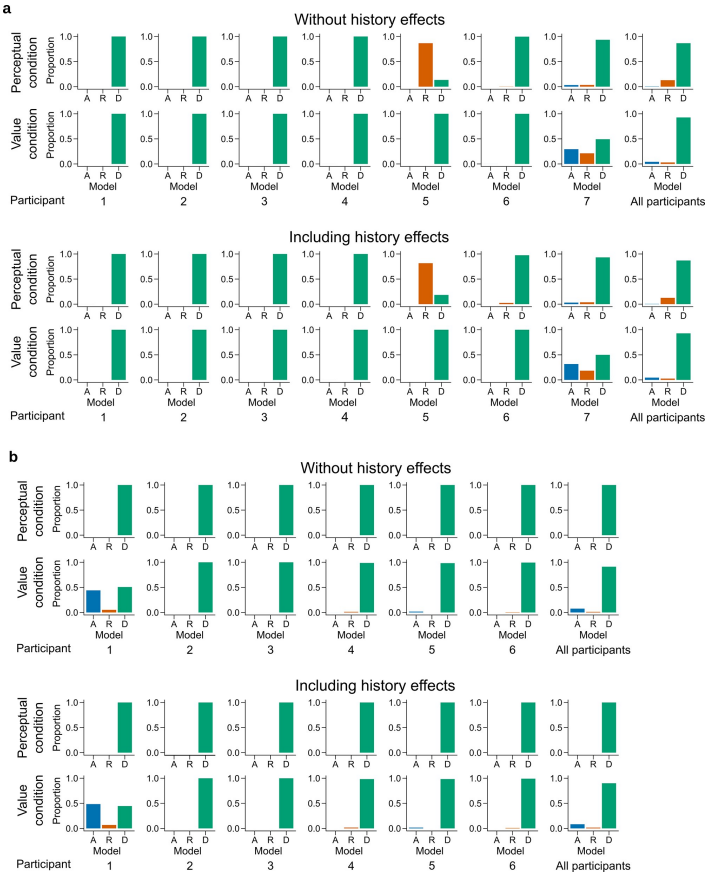


Figure 2.4 - figure supplement 1. Latent mixture model fits for each participant. Individual level fit of the latent mixture model excluding (top) or including (bottom) choice history effects for **(a)** Experiment 1 and **(b)** Experiment 2. The panels on the far right shows the average fit for all the participants of the given experiment. DbS is strongly favored for nearly all participants and clearly favored across participants, irrespective of the experimental condition. Including choice and correctness information of previous trials has minimal influence in the results of these analyses, which rules out the influence of these effects on the decision rule used by the participants.

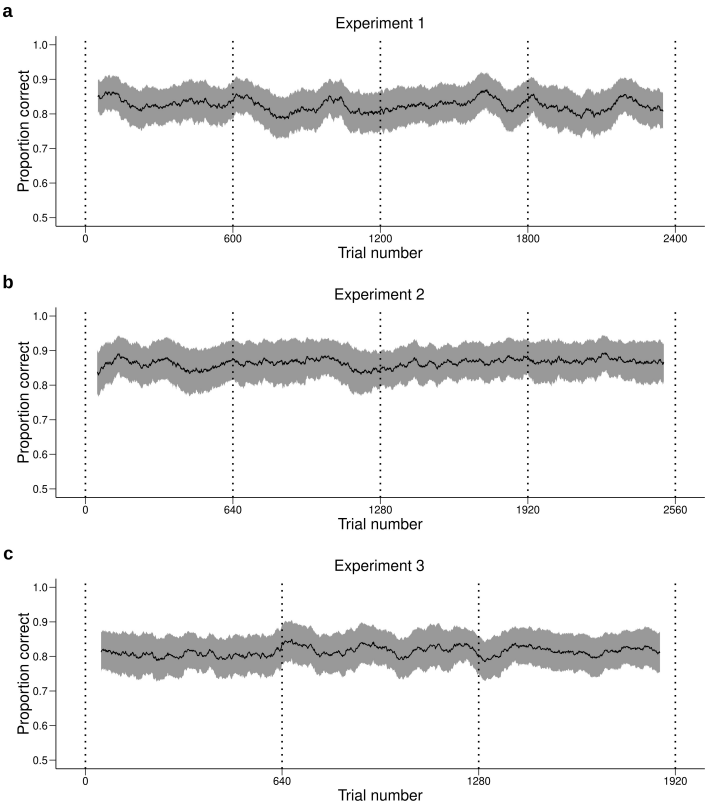


Figure 2.4 - figure supplement 2. Performance across time. Behavioral performance (mean ± SEM across participants) averaged over a moving window of 100 trials for **(a)** Experiment 1, **(b)** Experiment 2 and **(c)** Experiment 3. Each daily session took place between two dotted vertical lines. The performance of the participants is stable during and between daily sessions. Therefore, the quantitative and qualitative results presented in the main text are not likely to be influenced by changes in performance over time.

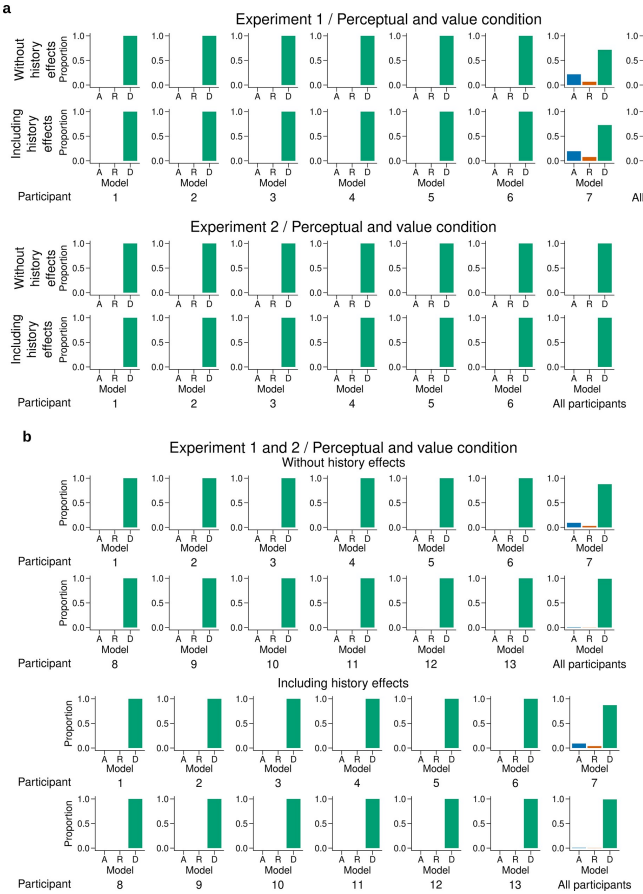


Figure 2.4 - figure supplement 3. Individual level fit of the latent mixture model combining data across experiments and experimental conditions. (a) Individual level fit of the latent mixture model combining data across both experimental conditions for Experiment 1 (top) and Experiment 2 (bottom). **(b)** Individual level fit of the latent mixture model combining data across both experimental conditions and both experiments. Each panel shows the results excluding (top) or including (bottom) choice history effects. The panels labeled 'All participants' show the average fit for all the participants of the given experiment. DbS is strongly favored irrespective of incentivized goals. Including the previous trial effects has minimal influence on these results.

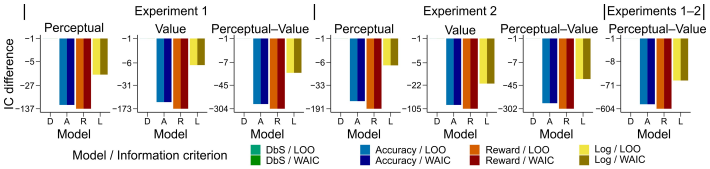


Figure 2.4 - figure supplement 4. Model comparison based on leave-one-out cross-validation metrics. Quantitative comparison of the models including choice and correctness effects of previous trials based on leave-one-out cross-validation metrics. Difference in LOO and WAIC between the best model (DbS (D) in all cases) and the competing models: Accuracy (A), Reward (R) and Logarithmic (L) models. Each panel shows the data grouped for each and across experiments and experimental conditions (see titles on top of each panel). Including the previous choice and correctness effects has only little influence on the results (compare with Figure 2.4b in main text). The DbS model provides the best fit to the behavioral data.

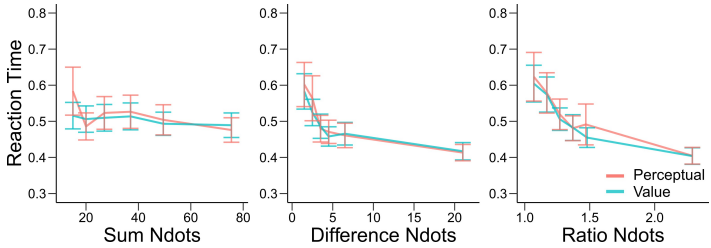


Figure 2.4 - figure supplement 5. Reaction times are similar in the perceptual and value conditions. Mean reaction times of participants in Experiments 1 and 2 in the perceptual (red) and value (blue) condition. Error bars represent SEM across participants. Reaction times are presented as a function of the sum of the number of dots in both clouds (left), the absolute difference between the number of dots in both clouds (middle) and the ratio of the number of dots in the most numerous cloud over the less numerous cloud (right). Non-parametric ANOVA tests revealed no significant differences in any of these behavioral assessments (all tests p>0.4).

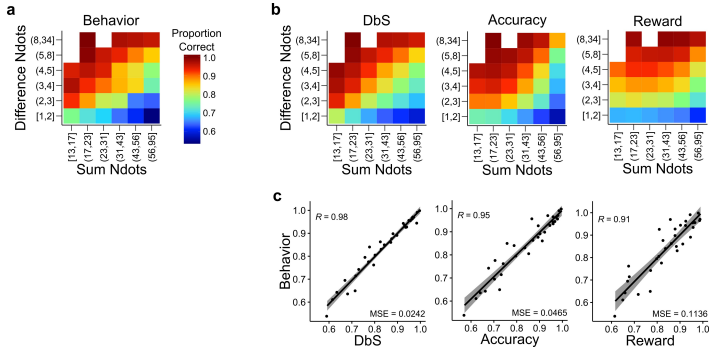


Figure 2.4 - figure supplement 6. Behavior and model predictions as a function of sum and difference in dots. (a) Average behavior in both conditions of Experiments 1 and 2 as a function of the sum of the number of dots in both clouds (Sum Ndots) and the absolute difference between the number of dots in both clouds (Difference Ndots). The data are binned as in Figure 2.4 but now expanded in two dimensions. **(b)** Predictions of each encoding rule model fit with only n as a free parameter shown with the same scale as in a. **(c)** Linear regression between the behavior for each combination of Sum Ndots and Difference Ndots bins and the predictions of each model for the same bins. DbS captures best the changes in behavior across bins of sum and absolute difference of the number of dots in both clouds. This analysis should not be considered as a quantitative proof, but as a qualitative inspection of the results presented in Figure 2.4.

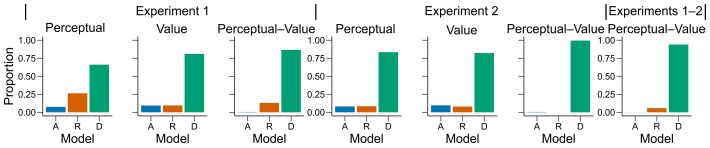


Figure 2.4 - figure supplement 7. Model fit for the first experimental condition of each participant. Similar as in Figure 2.4a, bars represent proportion of times an encoding rule (Accuracy [A, blue], Reward [R, red], DbS [D, green]) was selected by the Bayesian latent-mixture model based on the posterior estimates across participants. Each panel shows the data grouped for each experiment and experimental conditions (see titles on top of each panel). The latent-mixture model was only fit to the first condition that was carried out by each participant. As the participants did not know of the second condition before carrying it out, they could not adopt compromise strategies between the two objectives. Therefore, the fact that DbS is favored in the results is not an artifact of carrying out two different conditions in the same participants.

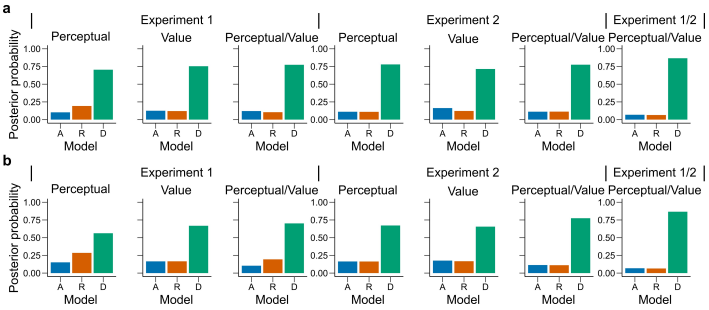


Figure 2.4 - figure supplement 8. Latent vector π posterior estimates. Bars represent the posterior distribution of the latent vector π, with each bar representing an encoding rule (Accuracy (A, blue), Reward (R, red), DbS (D, green)). Results are presented for **(a)** all sessions and **(b)** only the first condition carried out by each participant. DbS is consistently the most likely encoding rule.

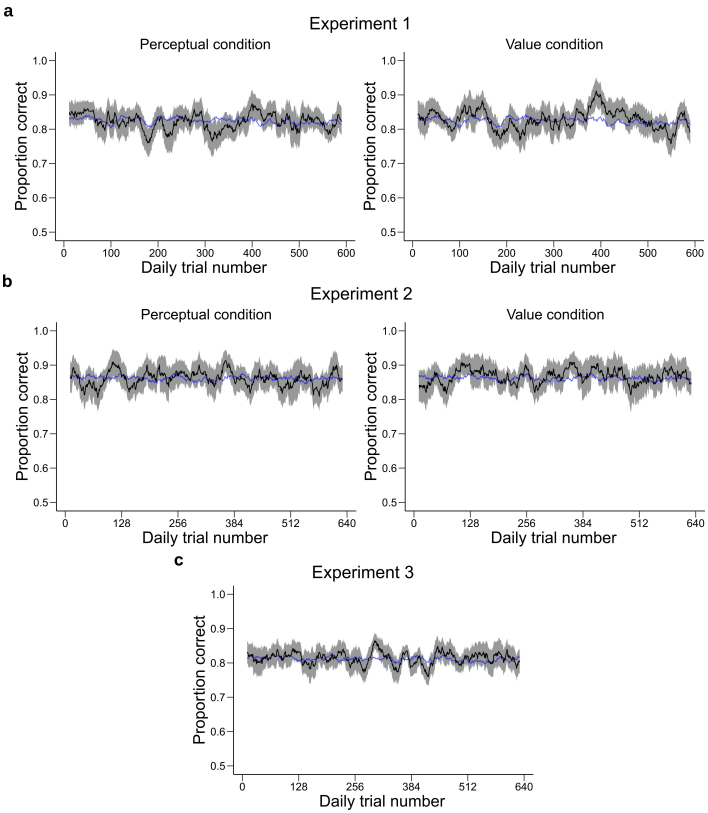


Figure 2.5 - figure supplement 1. Performance across trial experience. These plots represent the performance of the participants as a function of the number of trials they have experienced during the session. The performance of the participants (black, shaded area represents ± SEM across participants) was averaged over a moving window of 21 trials and is shown for Experiment 1 **(a)** Experiment 2 **(b)** and Experiment 3 **(c)**. The blue line represents the performance predicted by the α-adaptation model using the same moving window average. The model provides a good fit to average performance.

2.6. Appendix

Appendix 2.1: Infomax coding rule

We assume that the subjective perception of an environmental variable with value v is determined by n independent *samples* of a binary random variable, i.e. outcomes are either "high" (ones) or "low" (zeros) readings. Here, the probability θ of a “high” reading is the same on each draw, but can depend

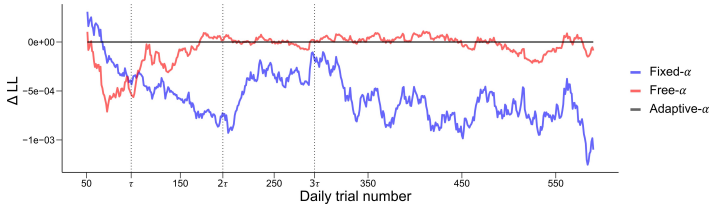


Figure 2.5 - figure supplement 2. Quantitative and dynamical analysis of adaptation over time. To further investigate the adaptation of the prior, we fit three models of varying complexity to the data of Experiments 1, 2 and 3. The Fixed-α model (blue) is defined with a fixed $\alpha = 2$. The Free-α model (red) allows the α parameter to vary across participants but is kept constant across time. The Adaptative-α corresponds to the model presented in Figure 2.5 where the prior adapts as the participants gains experience with the experimental distribution of dots. To allow a fair comparison with the Free-α model, the δ parameter, corresponding to the asymptotic value of the prior, was free to vary across participants. The log-likelihood of each model on each trial were averaged over a moving window of 100 trials and the log-likelihood of the Adaptative-α model was subtracted for comparison. Vertical dashed lines represent 1, 2 and 3 times τ, where τ controls the rate of adaptation in the Adaptative-α model. The Adaptative-α model provides a better fit for the first trials (until around 2τ), these trials correspond to the adaptation period where the α parameter is changing in the Adaptative-α model (see Figure 2.5). After this point the Adaptative-α and Free-α models provide a similar fit. This is to be expected as the function controlling the decay of α reaches its asymptotic value, leaving the two model virtually identical. The Fixed-α provides overall a worse fit, except for the early trials.

on the input stimulus value, via the function $\theta(v)$. Additionally, we assume that the input value v on a given trial is an independent draw from some prior distribution $f(v)$ in a given environment or context (with $F(v)$ being the corresponding cumulative distribution function). As we mentioned before, the choice of θ (i.e. encoding of the input vale) depends on v. Now suppose that the mapping $\theta(v)$ (the encoding rule) is chosen so as to maximize the mutual information between the random variable v and the subjective value representation k. The mutual information is computed under the assumption that v is drawn from a particular prior distribution $f(v)$, and $\theta(v)$ is assumed to be optimized for this prior. The mutual information between v and k is defined as

$$I(v,k) = H(k) - H(k|v), \tag{2.21}$$

where the marginal entropy $H(k)$ quantifies the uncertainty of the marginal response distribution $P(k)$, and $H(k|v)$ is the average conditional entropy of k given v. The output distribution is given by

$$P(k) = \int_{v \in V} P(k|v) f(v) dv, \tag{2.22}$$

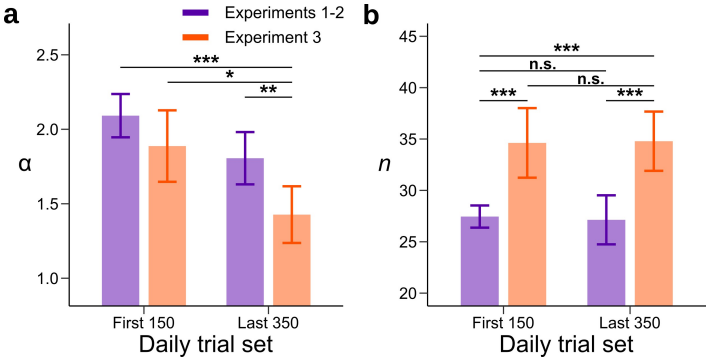


Figure 2.5 - figure supplement 3. Model fits for the beginning and end of each session without parametric assumptions. A model was fitted to the first 150 and last 350 trials of each daily session. The prior parameter α and the number of neural resources n were allowed to vary between the first and last sets of daily trials and between Experiments 1–2 (purple) and Experiment 3 (orange). **(a)** Each bar represents the mean value of the α parameter for a combination of experiments and set of daily trials. In Experiment 3, α is lower in the last set of trials compared to the first set of trials. In addition, the value of α for Experiment 3 is lower than for Experiments 1–2 in the last set of daily trials. **(b)** Each bar represents the value of the neural resource parameter n for a combination of experiments and set of daily trials. The neural resources parameter n in Experiment 3 is larger than in Experiments 1–2. However, there is no change in the neural resource parameter across the session. This suggests that the adaptation process is not an artifact of changes in the neural resource parameter, which could for example change with the engagement of the participants across the session. Significance between parameters was computed by subtracting the chain with the largest mean to the other one and measuring the proportion of values that fall below 0 (n.s. P $>$ 0.05, *P$<$0.05, **P$<$0.01, and ***P$<$0.001). Error bars represent $\pm$ SD of the posterior chains of the corresponding parameter.

where $f(v)$ is defined as the input density function. For the encoding framework that we consider here which is given by the binomial channel, the conditional probability mass function of the output given the input is

$$P(k|v) = \binom{n}{k}\theta(v)^k(1-\theta(v))^{n-k}, \qquad k \in [0,1,\ldots,n]. \tag{2.23}$$

Thus, we have all the ingredients to write the expression of the mutual information

$$\begin{aligned} I(v,k) &= H(k) - H(k|v) \\ &= -\sum_{k=0}^{n} P(k)\log P(k) \\ &\quad - \left(- \int_{v \in V} f(v) \sum_{k=0}^{n} P(k|v)\log P(k|v) \; dv \right) \end{aligned} \tag{2.24}$$

We then seek to determine the encoding rule $\theta(v)$ that solves the optimization problem

$$\text{find} \qquad C = \max_{\{\theta(v)\}} I(v,k). \tag{2.25}$$

It can be shown that for large n, the mutual information between θ and k (hence the mutual information between v and k) is maximized if the prior distribution over θ is the Jeffreys prior [91]

$$\text{Beta}(\theta; 0.5, 0.5) = \frac{1}{\pi\sqrt{\theta(1-\theta)}}, \tag{2.26}$$

also known as the arcsine distribution. Hence, the mapping $\theta(v)$ induces a prior distribution over θ given by the arcsine distribution. This means that for each v, the encoding function $\theta(v)$ must be such that

$$\begin{aligned} F(v) &= \int_0^{\theta(v)} \frac{1}{\pi\sqrt{\tilde{\theta}(1-\tilde{\theta})}} d\tilde{\theta} \\ &= \frac{2}{\pi}\text{arcsin}(\sqrt{\theta(v)}). \end{aligned} \tag{2.27}$$

Solving for θ we finally obtain the optimal encoding rule

$$\theta(v) = \left[\sin\left(\frac{\pi}{2}F(v)\right)\right]^2. \tag{2.28}$$

Appendix 2.2: Accuracy maximization for a known prior distribution

Here we derive the optimal encoding rule when the criterion to be maximized is the probability of a correct response in a binary comparison task, rather than mutual information as in Appendix 2.1. As in Appendix 2.1, we assume that the prior distribution $f(x)$ from which stimuli are drawn is known, and that the encoding rule is optimized for this particular distribution. (The case in which we wish the encoding rule to be robust to variations in the distribution from which stimuli are drawn is instead considered in Appendix 2.6.) Note that the objective assumed here corresponds to maximization of expected reward in the case of a perceptual experiment in which a subject must indicate which of two presented magnitudes is greater, and is rewarded for the number of correct responses. (In Appendix 2.5, we instead consider the encoding rule that would maximize expected reward if the subject's reward is proportional to the magnitude selected by their response.)

As above, we assume encoding by a binomial channel. The encoded value (number of "high" readings) is given by k, which is consequently an integer between 0 and n. This is a random variable with a binomial distribution with expected value and variance given by

$$\mathrm{E}\left[\frac{k}{n}|\theta\right] = \theta \qquad \mathrm{Var}\left[\frac{k}{n}|\theta\right] = \frac{\theta(1-\theta)}{n} \tag{2.29}$$

Suppose that the task of the decision maker is to decide which of two input values v_1 and v_2 is larger. Assuming that v_1 and v_2 are encoded independently, then the decision maker choses v_1 if and only if the internal readings $k_1 > k_2$ (here we may suppose that the probability of choosing stimulus 1 is 0.5 in the event that $k_1 = k_2$). Thus, the probability of choosing stimulus 1 is:

$$\mathrm{P}\left(\frac{k_1}{n} > \frac{k_2}{n}|v_1, v_2\right) + \frac{1}{2}\mathrm{P}\left(\frac{k_1}{n} = \frac{k_2}{n}|v_1, v_2\right). \tag{2.30}$$

In the case of large n, we can use a normal approximation to the binomial distribution to obtain

$$\left(\frac{k_1}{n} - \frac{k_2}{n}\right) \sim \mathrm{N}\left(\theta_1 - \theta_2, \frac{\theta_1(1-\theta_1) + \theta_2(1-\theta_2)}{n}\right) \tag{2.31}$$

and hence the probability of choosing v_1 is given by

$$\mathrm{P}_{\text{choose } v_1} \approx \Phi\left(\frac{\theta_1 - \theta_2}{\sqrt{\frac{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}{n}}}\right), \tag{2.32}$$

where $\Phi(\cdot)$ is the standard CDF. Thus the probability of an incorrect choice (i.e. choosing the item with the lower value) is approximately

$$\mathrm{P}_{\text{error}} \approx \Phi\left(-\frac{|\theta_1-\theta_2|}{\sqrt{\frac{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}{n}}}\right) \tag{2.33}$$

Now, suppose that the encoding rule, together with the prior distribution for v (the same for both inputs, that are independent draws from the prior distribution) results in an ex-ante distribution for θ (same for both goods) with density function $\hat{f}(\theta)$. Then the probability of error is given by

$$\mathrm{P}_{\text{error}} \approx \Phi\left(-\frac{|\theta_1-\theta_2|}{\sqrt{\frac{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}{n}}}\right) \hat{f}(\theta_1)\hat{f}(\theta_2)\, d\theta_1 d\theta_2 \tag{2.34}$$

Our goal is to evaluate Eq. 2.34 for any choice of the density $\hat{f}(\theta)$. First, we fix the value of θ_1 and integrate over θ_2:

$$\begin{aligned}
&{}_0^1\Phi\left(-\frac{|\theta_1-\theta_2|}{\sqrt{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}}\sqrt{n}\right)\hat{f}(\theta_2)\, d\theta_2 \\
&=_0^{\theta_1}\Phi\left(-\frac{\theta_2-\theta_1}{\sqrt{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}}\sqrt{n}\right)\hat{f}(\theta_2)\, d\theta_2 \\
&\quad +_{\theta_1}^1\Phi\left(-\frac{\theta_1-\theta_2}{\sqrt{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}}\sqrt{n}\right)\hat{f}(\theta_2)\, d\theta_2
\end{aligned} \tag{2.35}$$

with $\theta_2 = \theta_1 + \sqrt{2n\theta_1(1-\theta_1)}z$, the expression above then becomes

$$\begin{aligned}
&\approx_{\frac{-\theta_1\sqrt{n}}{\sqrt{2\theta_1(1-\theta_1)}}}^{0}\Phi(z)\hat{f}(\theta_1)\left[\frac{\sqrt{2\theta_1(1-\theta_1)}}{\sqrt{n}}\right]dz \\
&\quad +_0^{\frac{(1-\theta_1)\sqrt{n}}{\sqrt{2\theta_1(1-\theta_1)}}}\Phi(-z)\hat{f}(\theta_1)\left[\frac{\sqrt{2\theta_1(1-\theta_1)}}{\sqrt{n}}\right]dz \\
&\approx \underbrace{\left[2\int_{-\infty}^{0}\Phi(z)dz\right]}_{>0}\hat{f}(\theta_1)\frac{\sqrt{2\theta_1(1-\theta_1)}}{\sqrt{n}}
\end{aligned} \tag{2.36}$$

Then we can integrate over θ_1 to obtain:

$$\mathrm{P}_{\text{error}} \approx \frac{2}{\sqrt{n\pi}} \int \hat{f}(\theta_1)^2 \sqrt{(\theta_1(1-\theta_1))}\, d\theta_1. \tag{2.37}$$

This problem can be solved using the method of Lagrange multipliers:

$$\int \sqrt{\theta(1-\theta)}\hat{f}(\theta)^2 d\theta + \lambda(\int \hat{f}(\theta) - 1)$$

$$= \int (\sqrt{\theta(1-\theta)}\hat{f}(\theta)^2 + \lambda \hat{f}(\theta)) d\theta - \lambda$$

$$= \int \mathcal{L}(\theta, \hat{f}, \lambda)\, d\theta - \lambda \tag{2.38}$$

We now calculate the gradient

$$\frac{\partial \mathcal{L}}{\partial \hat{f}} = 2\hat{f}\sqrt{(\theta(1-\theta))} + \lambda \tag{2.39}$$

and then find the optimum for $\hat{f}$ by setting

$$2\hat{f}\sqrt{(\theta(1-\theta))} + \lambda = 0 \tag{2.40}$$

then solving for $\hat{f}$ to obtain

$$\hat{f} = \frac{-\lambda}{2\sqrt{\theta(1-\theta)}}. \tag{2.41}$$

Taken into consideration our optimization constraint, it can be shown that

$$\int_0^1 \frac{1}{\sqrt{\theta(1-\theta)}} = \frac{1}{\pi}$$

and therefore this implies:

$$\frac{1}{\pi} = \frac{-\lambda}{2}$$

thus requiring:

$$-\lambda = \frac{2}{\pi}.$$

Replacing λ in Eq. 2.41 we finally obtain

$$\hat{f}(\theta) = \frac{1}{\pi\sqrt{\theta(1-\theta)}} \tag{2.26 revisited}$$

Thus the optimal encoding rule is the same (at least in the large-n limit) in this case as when we assume an objective of maximum mutual information (the case considered in Appendix 2.1), though here we assume that the objective is accurate performance of a specific discrimination task.

Appendix 2.3: Optimal noise for a known prior distribution

Interestingly, we found that the fundamental principles of the theory independently developed in our work are directly linked to the concept of suprathreshold stochastic resonance (SSR) discovered about two decades ago. Briefly, SSR occurs in an array of n identical threshold non-linearities, each of which is subject to independently sampled random additive noise (neurons in main text). SSR should not be confused with the standard stochastic resonance (SR) phenomenon. In SR, the amplitude of the input signal is restricted to values smaller than the threshold for SR to occur. On the other hand, in SSR random draws from the distribution of input values can exist above threshold levels. Using the simplified implementational scheme proposed in our work, it can be shown that mutual information $I(v,k)$ can be also optimized by finding the optimal noise distribution. This is important as it provides a normative justification as for why sampling must be noisy in capacity-limited systems. Actually, SSR was initially motivated as a model of neural arrays such as those synapsing with hair cells in the inner ear, with the direct application of establishing the mechanisms by which information transmission can be optimized in the design of cochlear implants [135]. Our goal in this subsection is to make evident the link between the novel theoretical implications of our work and the SSR phenomenon developed in previous work [93, 135], which should further justify our argument of efficient noisy sampling as a general framework for decision behavior, crucially, with a parsimonious implementational nature.

Following our notation, each threshold device (we will call it from now on a *neuron*) can be seen as the number of n resources available to encode an input stimulus v. Here, we assume that each neuron produces a "high" reading if and only if $v+\eta>\tau$, where η is i.i.d. random additive noise (independent of v) following a distribution function f_η, and τ is the minimum threshold required to produce a "high" reading. If we define the noise CDF as F_η, then the probability θ of the neuron giving a "high" reading in response to the input signal v is given by

$$\theta(v) = 1 - F_\eta(\tau - v). \tag{2.42}$$

It can be shown that the mutual information between the input v and the number of "high" readings k for large n is given by [93]

$$I(v,k) \approx \frac{1}{2}\log_2\left(\frac{n\pi}{2e}\right) - D_{\mathrm{KL}}[f(v)||f_J(v)], \tag{2.43}$$

where f_J is the Jeffreys prior (Eq. 2.26). Therefore, Jeffreys' prior can also be derived making it a function of the noise distribution f_η

$$f_J(v) = \frac{f_\eta(\tau - v)}{\pi\sqrt{F_\eta(\tau - v)[1 - F_\eta(\tau - v)]}}. \tag{2.44}$$

Given that the first term in Eq. 2.43 is always non-negative, a sufficient condition for achieving channel capacity is given by

$$f(v) = f_J(v) \qquad \forall v. \tag{2.45}$$

Typically, the nervous system of any organism has little influence on the distribution of physical signals in the environment. However, it has the ability to shape its internal signals to optimize information transfer. Therefore, a parsimonious solution that the nervous system may adopt to adapt to statistical regularities of environmental signals in a given context is to find the optimal noise distribution f_η^* to achieve channel capacity. Note that this is different from classical problems in communication theory where the goal is usually to find the signal distribution that maximizes mutual information for a channel. Solving Eq. 2.44 to find $f_\eta(v)$ one can find such optimal noise distribution

$$f_\eta^*(v) = \frac{\pi}{2}\sin[\pi(1 - F(\tau - v))]f(\tau - v). \tag{2.46}$$

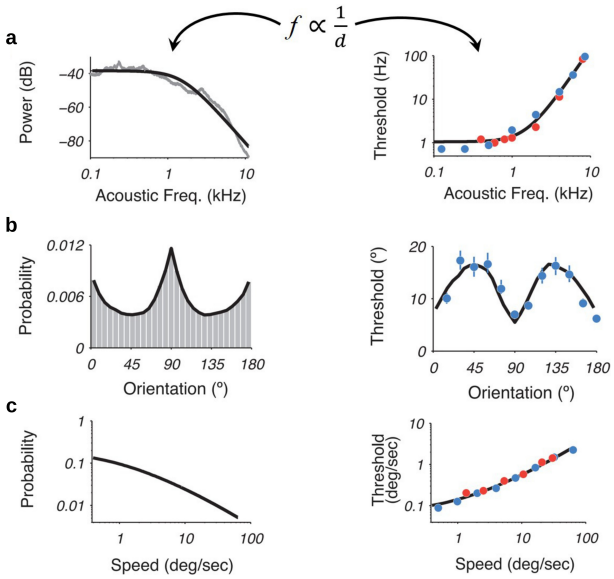
A further interesting consequence of this set of results is that the ratio between the signal PDF $f(v)$ and the noise PDF f_η is

$$\frac{f(v)}{f_\eta(\tau - v)} = \frac{2}{\pi\sin[\pi(1 - F(v))]}. \tag{2.47}$$

Using the definition given in Eq. 2.42 to make this expression a function of θ, one finds the optimal PDF of the encoder

$$f^*(\theta) = \frac{1}{\pi\sqrt{\theta(1-\theta)}}, \tag{2.48}$$

which is once again the arcsine distribution.



Appendix 2.4 - Figure 2.1. Recently, it was shown that using an efficiency principle for encoding sensory variables, based on population of noisy neurons, it was possible to obtain an explicit relationship between the statistical properties of the environment (the prior) and perceptual discriminability [25].The theoretical relation states that discriminability should be inversely proportional to the density of the prior distribution. Interestingly, this relationship holds across several sensory modalities such as **(a)** acoustic frequency, **(b)** local orientation, **(c)** speed (figure adapted with permission from the authors [25]). Here, we investigate whether this particular relation also emerges in our efficient sampling framework.

Appendix 2.4: Efficient coding and the relation between environmental priors and discrimination

We first show that we obtain a prediction of exactly the same kind from our model of encoding using a binary channel, in the case that (i) we assume that the encoding rule is optimized for a single environmental distribution, as in the theory of [25, 71], and (ii) the objective that is maximized is either mutual information (as in the theory of Ganguli and Simoncelli) or the probability of an accurate binary comparison (as considered in Appendix 2.2).

Note that the expected value and variance of a binomial random variable are given by

$$\mathrm{E}\left[r|\theta\right] = \theta \qquad \mathrm{Var}\left[r|\theta\right] = \frac{\theta(1-\theta)}{n}, \tag{2.49}$$

where we let here $r \equiv k/n$. In Appendix 2.2, we show that if the objective is accuracy maximization, an efficient binomial channel requires that

$$\theta(v) = \left[\sin\left(\frac{\pi}{2}F(v)\right)\right]^2.$$

Thus, replacing $\theta(v)$ in Eq. 2.49 implies the following relations

$$\mathrm{E}\left[r|\theta\right] = \sin^2(\omega), \qquad \mathrm{Var}\left[r|\theta\right] = \frac{\sin^2(\omega)\cos^2(\omega)}{n}, \tag{2.50}$$

where we let here $\omega \equiv \frac{\pi}{2}F(v)$. Discrimination thresholds d in sensory perception are defined as the ratio between the precision of the representation and the rate of change in the perceived stimulus

$$d \equiv \frac{\sqrt{\mathrm{Var}\left[r|\theta\right]}}{\mathrm{E}\left[r|\theta\right]'}. \tag{2.51}$$

Substituting the expressions for expected value and variance in Eq. 2.50 results in

$$\begin{aligned} d &= \frac{1}{2\sqrt{n}\omega'} \\ &= \frac{1}{\sqrt{n}\pi f(v)}. \end{aligned} \tag{2.52}$$

Thus under our theory, this implies

$$d \propto \frac{1}{f(v)}. \tag{2.53}$$

This is exactly the relationship derived and tested by [25].

Our model instead predicts a somewhat different relationship if the encoding rule is required to be robust to alternative possible environmental frequency distributions (the case further discussed in Appendix 2.6). In this case, the robustly optimal encoding rule is DbS, which corresponds to $\theta(v) = F(v)$, rather than the relation 2.53. Substituting this into Eqs. 2.49 and 2.51 yields the prediction

$$d \;=\; \frac{\sqrt{F(v)(1-F(v))}}{\sqrt{n}} \cdot \frac{1}{f(v)}. \tag{2.54}$$

instead of Eq. 2.52.

One interpretation of the experimental support for the relation 2.53 reviewed by [25] could be that in the case of early sensory processing of the kind with which they are concerned, perceptual processing is optimized for a particular environmental frequency distribution (representing the long-run experience of an organism or even of the species), so that the assumptions used in Appendix 2.2 are the empirically relevant ones. Even so, it is arguable that robustness to changing contextual frequency distributions should be important in the case of higher forms of cognition, so that one might expect prediction 2.54 to be more relevant for these cases; and indeed, our experimental results for the case of numerosity discrimination are more consistent with Eq. 2.54 than with 2.52.

One should also note that even in a case where Eq. 2.54 holds, if one measures discrimination thresholds over a subset of the stimulus space, over which there is non-trivial variation in $f(v)$, but $F(v)$ does not change very much (because the prior distribution for which the encoding rule is optimized assigns a great deal of probability to magnitudes both higher and lower than those in the experimental data set), then relation (2.54) restricted to this subset of the possible values for v will imply that the relation (2.53) should approximately hold. This provides another possible interpretation of the fact that the relation (2.53) holds fairly well in the data considered by [25].

Appendix 2.5: Maximizing expected size of the selected item (fitness maximization)

We now consider the optimal encoding rule under a different assumed objective, namely, maximizing the expected magnitude of the item selected by the subject's response (that is, the stimulus judged to be larger by the subject), rather than maximizing the probability of a correct response as in Appendix 2.2. While in many perceptual experiments, maximizing the probability of a correct response would correspond to maximization of the subject's expected reward (or at least maximization of a psychological reward to the subject, who is given feedback about the correctness of responses but not about true magnitudes), in many of the ecologically relevant cases in which accurate discrimination of numerosity is useful to an organism [48, 87], the decision maker's reward depends on how much larger one number is than another, and not simply their ordinal ranking. This would also be true of typical cases in which internal representations of numerical magnitudes must be used in economic decision making: the reward from choosing an investment with a larger monetary payoff is proportional to the size of the payoff afforded by the option that is chosen. Hence it is of interest to consider the optimal encoding rule if we suppose that encoding is optimized to maximize performance in a decision task with this kind of reward structure.

As in Appendix 2.1 and Appendix 2.2, we again consider the problem of optimizing the encoding rule for a specific prior distribution $f(v)$ for the magnitudes that may be encountered, and we assume that it is only possible to encode information via "high" or "low" readings. The optimization problem that we need to solve is to find the optimal encoding function $\theta(v)$ that guarantees a maximal expected value of the chosen outcome, for any given prior distribution $f(v)$. Thus the quantity that we seek to maximize is given by

$$\begin{aligned} \mathrm{E}[v(\text{chosen})] \;=\; \int\int f(v_1, v_2)\,[P_1(\theta(v_1),\, \theta(v_2))v_1 + \\ P_2(\theta(v_1),\, \theta(v_2))v_2]\,dv_1 dv_2 \end{aligned} \tag{2.55}$$

where $P_i(\theta_1, \theta_2)$ is the probability of choosing option i when the encoded values of the two options are θ_1 and θ_2 respectively.

We begin by noting that for any pair of input values v_1, v_2, the integrand in (2.55) can be written as

$$\begin{aligned} &P_1(\theta(v_1),\, \theta(v_2))v_1 \,+\, P_2(\theta(v_1),\, \theta(v_2))v_2 \\ &\quad = \max(v_1, v_2) \,-\, P(\text{error}\,|\theta(v_1), \theta(v_2))\,|v_1 - v_2|, \end{aligned} \tag{2.56}$$

where $I(A)$ is the indicator function (taking the value 1 if statement A is true, and the value 0 otherwise), and $P(\text{error}\,|\theta_1, \theta_2)$ is the probability of choosing the lower-valued of the two options.

Substituting this last expression for the integrand in (2.55), we see that we can equivalently write

$$\begin{aligned} \text{E}[v(\text{chosen})] \;&=\; \text{E}[\max(v_1, v_2)] - \\ \int\int f(v_1, v_2)\, P(\text{error}\,|\theta(v_1), \theta(v_2))\,|v_1 - v_2|\, dv_1 dv_2, \end{aligned} \tag{2.57}$$

where

$$\text{E}[\max(v_1, v_2)] \;\equiv\; \int\int f(v_1, v_2)\, \max(v_1, v_2)\, dv_1 dv_2 \tag{2.58}$$

is a quantity which is independent of the encoding function $\theta(v)$. Hence choosing $\theta(v)$ to maximize (2.55) is equivalent to choosing it to minimize

$$\text{E}[\text{loss}] \;=\; \int\int f(v_1, v_2)\, P(\text{error}\,|\theta(v_1), \theta(v_2))\,|v_1 - v_2|\, dv_1 dv_2. \tag{2.59}$$

As previously specified, the probability of error given two internal noisy readings k_1 and k_2 is given by

$$P(\text{error}) = \left(\frac{k_1}{n} - \frac{k_2}{n} > 0 | v_1, v_2\right) \tag{2.60}$$

$$\approx \Phi\left(\frac{\theta_1 - \theta_2}{\sqrt{\frac{\theta_1(1-\theta_1)+\theta_2(1-\theta_2)}{n}}}\right), \tag{2.61}$$

where in this case we assume that v_1 is the lower-valued option and v_2 is the higher-valued option on any given trial. This implies that $P(\text{error})$ is very close to zero, except when $|\theta_1 - \theta_2| = \mathcal{O}(1/\sqrt{n})$. In this case we have

$$P(\text{error}) \approx \Phi\left(\sqrt{\frac{n}{2}}\frac{\theta_1 - \theta_2}{\sqrt{\theta(1-\theta)}}\right) \qquad \text{where} \quad \theta \equiv \frac{\theta_1 + \theta_2}{2}. \tag{2.62}$$

As in the case of accuracy maximization, here we assume that (v_1, v_2) are independent draws from the same distribution of possible values $f(v)$. Thus $f(v_1, v_2) = f(v_1)f(v_2)$. Then fixing v_1 and integrating over all possible values of v_2 in Eq. 2.59, the expected loss is approximately

$$\mathrm{E}[\text{loss}|v_1] = \int f(v_2)P(\text{error}|v_2, v_1)|v_2 - v_1|\, dv_2 \tag{2.63}$$

$$\approx \int f(v_2)\Phi\left(-\sqrt{\frac{n}{2}}\frac{|\theta_1 - \theta_2|}{\sqrt{\theta_1(1-\theta_1)}}\right)|v_2 - v_1|\, dv_2 \tag{2.64}$$

$$\approx f(v_1)\int \Phi\left(-\sqrt{\frac{n}{2}}\frac{\theta'(v_1)|v_2 - v_1|}{\sqrt{\theta_1(1-\theta_1)}}\right)|v_2 - v_1|\, dv_2 \tag{2.65}$$

$$\approx f(v_1)\int_{-\infty}^{\infty}\Phi(-|z|)\left[\sqrt{\frac{2}{n}}\frac{\theta_1(1-\theta_1)}{\theta'(v_1)}|z|\right]\left[\sqrt{\frac{2}{n}}\frac{\theta_1(1-\theta_1)}{\theta'(v_1)}\right] dz \tag{2.66}$$

$$\approx \frac{4}{n}\frac{f(v_1)}{\theta'(v_1)^2}[\theta_1(1-\theta_1)]\underbrace{\int_0^{\infty}\Phi(-z)z\, dz}_{1/4} \tag{2.67}$$

$$\approx \frac{1}{n}\frac{f(v_1)}{\theta'(v_1)^2}[\theta_1(1-\theta_1)] \tag{2.68}$$

where in Eq. 2.66 we have applied the change of variable

$$z \equiv \frac{n}{2}\frac{\theta'(v_1)}{\theta_1(1-\theta_1)}(v_2 - v_1) \tag{2.69}$$

and in the integral of Eq. 2.67 we have used

$$\int_0^{\infty}\Phi(-z)z\, dz = \frac{1}{2}\left[(z^2-1)\Phi(-z) - z\phi(-z)\right]_0^{\infty} \tag{2.70}$$

$$= \frac{1}{2}\left[0 - (-\frac{1}{2})\right] \tag{2.71}$$

$$= \frac{1}{4} \tag{2.72}$$

where $\phi()$ is the standard normal PDF. Then integrating over v_1, we have:

$$\mathrm{E}[\text{loss}] = \frac{1}{n}\int \frac{f(v_1)^2}{\theta'(v_1)^2}[\theta_1(1-\theta_1)]\, dv_1. \tag{2.73}$$

Thus we want to find the encoding rule $\theta(v)$ to minimize this integral given the prior $f(v)$. We now apply the change of variable $\theta(v) \equiv \sin^2(\gamma(v))$, where $\gamma(v)$ is an increasing function with a range $0 \leqslant \gamma(v) \leqslant \frac{\pi}{2}$ for all v. Then we have

$$\theta'(v) = 2\sin(\gamma(v))\cos(\gamma(v))\gamma'(v) \tag{2.74}$$

$$= 2\sqrt{\theta(v)(1-\theta(v))}\gamma'(v) \tag{2.75}$$

and therefore we have

$$\frac{\theta(v)(1-\theta(v))}{\theta'(v)} = \frac{1}{4}\frac{1}{\gamma'(v)}. \tag{2.76}$$

This allows us to rewrite Eq. 2.73 as follows

$$\mathrm{E}[\text{loss}] = \frac{1}{n}\int \frac{f(v)^2}{\gamma'(v)^2}. \tag{2.77}$$

Now the problem is to choose the function $\gamma(v)$ to minimize E[loss] subject to $0 \leqslant \gamma(v) \leqslant \frac{\pi}{2}$. Equivalently, we can choose the function $\gamma'(v) > 0$ to minimize E[loss] subject to $\int \gamma'(v)dv \leqslant \frac{\pi}{2}$. Defining $\varphi(v) \equiv \gamma'(v)$, the optimization problem to solve is to choose the function $\varphi(v)$ to

$$\min \quad \int \frac{f(v)^2}{\varphi(v)^2}dv \qquad \text{s.t.} \quad \int \varphi(v)\,dv \leqslant \frac{\pi}{2} \tag{2.78}$$

Due to FOC, it can be shown that

$$\frac{f(v)^2}{\varphi(v)^3} \;=\; \text{same for all } v \quad \Rightarrow \quad \varphi(v) \sim f(v)^{2/3}. \tag{2.79}$$

Note also that the constraint $\int \varphi(v) \leqslant \frac{\pi}{2}$ must hold with equality, thus arriving at

$$\gamma(v) = \frac{\pi}{2}\int_{-\infty}^{v} f(\tilde{v})^{2/3}\,d\tilde{v}\int_{-\infty}^{\infty} f(\tilde{v})^{2/3}\,d\tilde{v}. \tag{2.80}$$

Therefore, we finally obtain the efficient encoding rule that maximizes the expected magnitude of the selected item

$$\theta(v) = \sin\left[\frac{\pi}{2}\int_{-\infty}^{v} f(\tilde{v})^{2/3}\,d\tilde{v}\int_{-\infty}^{\infty} f(\tilde{v})^{2/3}\,d\tilde{v}\right]^2 \tag{2.81}$$

Appendix 2.6: Robust optimality of DbS among encoding rules with m=1

Here we consider the nature of the optimal encoding function when the cost of increasing the size of the sample of values from prior experience that are used to adjust the encoding rule to the contextual distribution of stimulus values is great enough to make it optimal to base the encoding of a new stimulus magnitude v on a single sampled value $\tilde{v}$ from the contextual distribution. (The conditions required for this to be the case are discussed further in Appendix 2.7)

We assume that for each of the n independent processing units, the probability of a "high" reading is given by $\theta(v, \tilde{v}_j)$, where $\tilde{v}_j$ is the draw from the contextual distribution by processor j, and $\theta(v, \tilde{v})$ is the same function for each of the processing units. The $\{\tilde{v}_j\}$ for $j = 1, 2, \ldots, n$, are independent draws from the contextual distribution $f(v)$. We further assume that the function $\theta(v, \tilde{v})$ satisfies certain regularity conditions. First, we assume that θ is a piecewise continuous function. That is, we assume that the $v - \tilde{v}$ plane can be divided into a countable number of connected regions, with the boundaries between regions defined by continuous curves; and that the function $\theta(v, \tilde{v})$ is continuous in the interior of any of these regions, though it may be discontinuous at the boundaries between regions. And second, we assume that $\theta(v, \tilde{v})$ is necessarily weakly increasing in v and weakly decreasing in $\tilde{v}$. The function is otherwise unrestricted.

For any prior distribution $f(v)$ and any encoding function $\theta(v, \tilde{v})$, we can compute the probability of an erroneous comparison when two stimulus magnitudes v_1, v_2 are independently drawn from the distribution $f(v)$, and each of these stimuli is encoded using n additional independent draws $\{\tilde{v}_j\}$ from the same distribution. Let this error probability be denoted $P_n(\theta; f)$. We wish to find an encoding rule (for given n) that will make this error probability as small as possible; however, the answer to this question will depend on the prior distribution $f(v)$. Hence we wish to find an encoding rule that is *robustly* optimal, in the sense that it achieves the minimum possible value for the upper bound

$$\bar{P}_{error}(\theta) \equiv \sup_{f \in \mathcal{F}} P_n(\theta; f)$$

for the probability of an erroneous comparison. Here the class of possible priors $\mathcal{F}$ to considered is the set of all possible probability distributions (over values of v) that can be characterized by an integrable probability density function $f(v)$. (We exclude from consideration priors in which there is an

atom of probability mass at some single magnitude v, since in that case there would be a positive probability of a situation in which it is not clear which response should be considered "correct", so that P_{error} is not well-defined.) Note that the criterion $\bar{P}_{error}(\theta)$ for ranking encoding rules is not without content, since there exist encoding rules (including DbS) for which the upper bound is less than 1/2 (the error probability in the case of a completely uninformative internal representation).

Let us consider first the case in which there is some part of the diagonal line along which $\tilde{v} = v$ which is not a boundary at which the function $\theta(v, \tilde{v})$ is discontinuous. Then we can choose an open interval (v_{min}, v_{max}) such that all values $v, \tilde{v}$ with the property that both v and $\tilde{v}$ lie within the interval (v_{min}, v_{max}) are part of a single region on which $\theta(v, \tilde{v})$ is a continuous function. Then let θ_{min} be the greatest lower bound with the property that $\theta(v, \tilde{v}) \geq \theta_{min}$ for all $v, \tilde{v}$ lying within the specified interval, and similarly let θ_{max} be the lowest upper bound such that $\theta(v, \tilde{v}) \leq \theta_{max}$ for all values within the specified interval. Because of the continuity of $\theta(v, \tilde{v})$ on this region, as the values v_{min}, v_{max} are chosen to be close enough to each other, the bounds $\theta_{min}, \theta_{max}$ can be made arbitrarily close to one another.

Now for any probabilities $0 \leq \theta \leq \theta' \leq 1$, let $P_{min}(\theta, \theta')$ be the quantity defined in Eq. 2.30, when $\theta_1 = \theta$ and $\theta_2 = \theta'$; that is, for any v_1, v_2 that are not equal to one another, $P_{min}(\theta, \theta')$ is the probability of an erroneous comparison if the units representing the smaller magnitude each give a "high" reading with probability θ and those representing the larger magnitude each give a "high" reading with probability θ'. Then the probability of erroneous choice P_{error} when $f(v)$ is a distribution with support entirely within the interval (v_{min}, v_{max}) is necessarily greater than or equal to the lower bound $P_{min}(\theta_{min}, \theta_{max})$. The reason is that for any v_1, v_2 in the support of $f(v)$, the probabilities

$$\theta_i \;=\; \int \theta(v_i, \tilde{v}) f(\tilde{v}) d\tilde{v}$$

will necessarily lie within the bounds $\theta_{min} \leq \theta_i \leq \theta_{max}$ for both $i = 1, 2$. Given these bounds, the most favorable case for accurate discrimination between the two magnitudes will be to assign the largest possible probability θ_{max} to units being on in the representation of the larger magnitude, and the smallest possible probability θ_{min} to units being on in the representation of the smaller magnitude. Since the lower bound $P_{min}(\theta_{min}, \theta_{max})$ applies in the case of any individual values v_1, v_2 drawn from the support of $f(v)$, this same quantity is also a lower bound for the average error rate integrating over the prior distributions for v_1 and v_2.

One can also show that as the two bounds $\theta_{min}, \theta_{max}$ approach one another, the lower bound $P_{min}(\theta_{min}, \theta_{max})$ approaches $1/2$, regardless of the common value that θ_{min} and θ_{max} both approach. Hence it is possible to make $P_{min}(\theta_{min}, \theta_{max})$ arbitrarily close to $1/2$, by choosing values for v_{min}, v_{max} that are close enough to one another. It follows that for any bound P_{min} less than $1/2$ (including values arbitrarily close to $1/2$), we can choose a prior distribution $f(v)$ for which P_{error} is necessarily equal to P_{min} or larger. It follows that in the case of a function $\theta(v, \tilde{v})$ of this kind, the upper bound $\bar{P}_{error}(\theta)$ is equal to $1/2$.

In order to achieve an upper bound lower than $1/2$, then, we must choose a function $\theta(v, \tilde{v})$ that is discontinuous along the entire line $v = \tilde{v}$. For any such function, let us consider a value v^* with the property that all points $(v, \tilde{v})$ near (v^*, v^*) with $v > \tilde{v}$ belong to one region on which θ is continuous, and all points near (v^*, v^*) with $v < \tilde{v}$ belong to another region. Then under the assumption of piecewise continuity, $\theta(v, \tilde{v})$ must approach some value $\bar{\theta}(v^*)$ as the values $(v, \tilde{v})$ converge to (v^*, v^*) from within the region where $v > \tilde{v}$, and similarly $\theta(v, \tilde{v})$ must approach some value $\underline{\theta}(v^*)$ as the values $(v, \tilde{v})$ converge to (v^*, v^*) from within the region where $v < \tilde{v}$.

It must also be possible to choose values $v_{min} < v^* < v_{max}$ such that all points (v, v) with $v_{min} < v < v_{max}$ are points on the boundary between the two regions on which θ is continuous. Given such values, we can then define bounds $\underline{\theta}_{min}$, $\underline{\theta}_{max}$, $\bar{\theta}_{min}$, and $\bar{\theta}_{max}$, such that

$$\underline{\theta}_{min} \;\leq\; \theta(v, \tilde{v}) \;\leq\; \underline{\theta}_{max}$$

for all $v_{min} < v < \tilde{v} < v_{max}$, and

$$\bar{\theta}_{min} \;\leq\; \theta(v, \tilde{v}) \;\leq\; \bar{\theta}_{max}$$

for all $v_{min} < \tilde{v} < v < v_{max}$. Moreover, piecewise continuity of the function $\theta(v, \tilde{v})$ implies that by choosing both v_{min} and v_{max} close enough to v^* we can make the bounds $\underline{\theta}_{min}, \underline{\theta}_{max}$ arbitrarily close to $\underline{\theta}(v^*)$, and make the bounds $\bar{\theta}_{min}, \bar{\theta}_{max}$ arbitrarily close to $\bar{\theta}(v^*)$.

Next, for any set of four probabilities $0 \leq \underline{\theta} \leq \underline{\theta}' \leq 1$ and $0 \leq \bar{\theta} \leq \bar{\theta}' \leq 1$, let us define

$$\hat{P}_{min}(\underline{\theta}, \underline{\theta}'; \bar{\theta}, \bar{\theta}') \;\equiv\; \mathrm{E}[P_{min}(\theta(z_1), \theta'(z_2)) \,|\, z_1 < z_2], \tag{2.82}$$

where

$$\theta(z) \;\equiv\; z\bar{\theta} + (1-z)\underline{\theta}, \quad \theta'(z) \;\equiv\; z\bar{\theta}' + (1-z)\underline{\theta}', \tag{2.83}$$

and z_1, z_2 are two independent random variables, each distributed uniformly on [0, 1]. Then if $\theta(v, \tilde{v})$ lies between the lower bound $\underline{\theta}$ and upper bound $\underline{\theta}'$ whenever $v < \tilde{v}$, and between the lower bound $\bar{\theta}$ and upper bound $\bar{\theta}'$ whenever $v > \tilde{v}$, then the probability θ of a processing unit representing the magnitude v giving a "high" reading will lie between the bounds $\theta(z) \leq \theta \leq \theta'(z)$, where $z = F(v)$ is the quantile of v within the prior distribution. It follows that in the case of any two magnitudes v_1, v_2 with $v_1 < v_2$, the probability of an erroneous comparison will be bounded below by $P_{min}(\theta(z_1), \theta'(z_2))$, where $z_i = F(v_i)$ for $i = 1, 2$, since the probability of a correct discrimination will be maximized by making the units representing v_1 give as few high readings as possible and the units representing v_2 give as many high readings as possible. Integrating over all possible draws of v_1, v_2, one finds that the quantity $\hat{P}_{min}(\underline{\theta}, \underline{\theta}'; \bar{\theta}, \bar{\theta}')$ defined in (2.82) is a lower bound for the overall probability of an erroneous comparison, given that regardless of the prior $f(v)$, the quantiles z_1, z_2 will be two independent draws from the uniform distribution on $[0, 1]$.

Now consider again an encoding function $\theta(v, \tilde{v})$ of the kind discussed two paragraphs above, and an interval of stimulus values (v_{min}, v_{max}) of the kind discussed there. For any prior distribution $f(v)$ with support entirely contained within the interval (v_{min}, v_{max}), the probability of an erroneous comparison is bounded below by

$$P_n(\theta; f) \;\geq\; \hat{P}_{min}(\underline{\theta}_{min}, \underline{\theta}_{max};\, \bar{\theta}_{min}, \bar{\theta}_{max}),$$

where the function $\hat{P}_{min}$ is defined in (2.82). Moreover, by choosing the values v_{min}, v_{max} close enough to v^*, we can make this lower bound arbitrarily close to $P^e(\underline{\theta}(v^*), \bar{\theta}(v^*))$, where for any probabilities $\underline{\theta}, \bar{\theta}$ we define

$$P^e(\underline{\theta}, \bar{\theta}) \;\equiv\; \hat{P}_{min}(\underline{\theta}, \underline{\theta};\, \bar{\theta}, \bar{\theta}). \tag{2.84}$$

Hence in the case of the encoding function considered, the upper bound $\bar{P}_{error}(\theta)$ must be at least as large as $P^e(\underline{\theta}(v^*), \bar{\theta}(v^*))$. We further observe that the quantity $P^e(\underline{\theta}, \bar{\theta})$ defined in (2.84) is just the probability of an erroneous comparison in the case of an encoding rule according to which

$$\theta(v, \tilde{v}) \;=\; \underline{\theta} \qquad \text{if } v < \tilde{v},$$

$$\theta(v, \tilde{v}) \;=\; \bar{\theta} \qquad \text{if } v > \tilde{v}.$$

Note that in the case of such an encoding rule, the probability of an erroneous comparison is the same for all prior distributions, since under this rule all

that matters is the distribution of the quantile ranks of v and $\tilde{v}$. It is moreover clear that $P^e(\underline{\theta}, \bar{\theta})$ is an increasing function of $\underline{\theta}$ and a decreasing function of $\bar{\theta}$. It thus achieves its minimum possible value if and only if $\underline{\theta} = 0$ and $\bar{\theta} = 1$, in which case it takes the value P^{DbS}_{error}, the probability of erroneous comparison in the case of decision by sampling (again, independent of the prior distribution).

Thus in the case that there exists any magnitude v^* for which $\underline{\theta}(v^*) > 0$, $\bar{\theta}(v^*) < 1$, or both, there exist priors $f(v)$ for which $P_n(\theta; f)$ must exceed $P^{DbS}_{error} = P^e(0, 1)$. Hence in order to minimize the upper bound $\bar{P}_{error}(\theta)$, it must be the case that $\underline{\theta}(v) = 0$ and $\bar{\theta}(v) = 1$ for all v. But then our assumption that the encoding rule $\theta(v, \tilde{v})$ is at least weakly increasing in v and at least weakly decreasing in $\tilde{v}$ requires that

$$\theta(v, \tilde{v}) = 0 \qquad \text{for all } v < \tilde{v},$$

$$\theta(v, \tilde{v}) = 1 \qquad \text{for all } v > \tilde{v}.$$

Thus the encoding rule must be the DbS rule, the unique rule for which $\bar{P}_{error}(\theta)$ is no greater than P^{DbS}_{error}.

Appendix 2.7: Sufficient conditions for the optimality of DbS

Here we consider the general problem of choosing a value of m (the number of samples from the contextual distribution $f(v)$ to use in encoding any individual stimulus) and an encoding rule $\theta(v; \tilde{v}_1, \ldots, \tilde{v}_m)$ to be used by each of the n processing units that encode the magnitude of that single stimulus, so as to minimize the compound objective

$$\bar{P}_{error}(\theta) \; + \; K(m),$$

where $\bar{P}_{error}$ is the upper bound on the probability of an erroneous comparison under the encoding rule θ, and $K(m)$ is the cost of using a sample of size m when encoding each stimulus magnitude. The value of n is taken as fixed at some finite value. (This too can be optimized subject to some cost of additional processing units, but we omit formal analysis of this problem.) We assume that $K(m)$ is an increasing function of m, and can without loss of generality assume the normalization $K(0) = 0$. In this optimization problem, we assume that the only encoding functions θ to be considered are ones that are piecewise continuous, at least weakly increasing in v, and weakly decreasing in each of the $\tilde{v}_j$.

For any value of m, let $P^*(m)$ be the minimum achievable value for $\bar{P}_{error}(\theta)$. (Appendix 2.6 illustrates how this kind of problem can be solved, for the case $m = 1$.) Then the optimal value of m will be the one that minimizes $P^*(m) + K(m)$.

We can establish a lower bound for $P^*(m)$ that holds for any m:

$$\begin{aligned} P^*(m) \;\; &\equiv \;\; \inf_{\theta(v;\tilde{v}_1,\ldots,\tilde{v}_m)} \sup_{f\in\mathcal{F}} P_n(\theta; f) \\ &\geq \;\; \sup_{f\in\mathcal{F}} \inf_{\theta(v;\tilde{v}_1,\ldots,\tilde{v}_m)} P_n(\theta; f) \\ &= \;\; \sup_{f\in\mathcal{F}} \inf_{\theta(v)} P_n(\theta; f) \;\equiv\; \underline{P}_n. \end{aligned} \tag{2.85}$$

In the second line, we allow the function $\theta(v; \tilde{v}_1, \ldots, \tilde{v}_m)$ to be chosen after a particular prior $f(v)$ has already been selected, which cannot increase the worst-case error probability. In the third line, we note that the only thing that matters about the encoding function chosen in the second line is the mean value of $\theta(v; \tilde{v}_1, \ldots, \tilde{v}_m)$ for each possible magnitude v, integrating over the possible samples of size m that may be drawn from the specified prior; hence we can more simply write the problem on the second line as one involving a direct choice of a function $\theta(v)$, which may be different depending on the

prior $f(v)$ that has been chosen. The problem on the third line defines a bound $\underline{P}_n$ that does not depend on m.

A set of sufficient conditions for $m = 1$ to be optimal is then given by the assumptions that

(a) $P^*(0) > P^*(1) + K(1)$, and

(b) $P^*(1) - \underline{P} < K(2) - K(1)$.

Condition (a) implies that $m = 0$ will be inferior to $m = 1$: the cost of a single sample is not so large as to outweigh the reduction in $\bar{P}_{error}(\theta)$ that can be achieved using even one sample. Condition (b) implies that $m = 1$ will be superior to any $m' > 1$. The lower bound (2.85), together with our monotonicity assumption regarding $K(m)$, implies that for any $m' > 1$,

$$P^*(1) - P^*(m') \;\leq\; P^*(1) - \underline{P} \;<\; K(2) - K(1) \;\leq\; K(m') - K(1),$$

and hence that

$$P^*(1) + K(1) \;<\; P^*(m') + K(m').$$

While condition (b) is stronger than is needed for this conclusion, the sufficient conditions stated in the previous paragraph have the advantage that we need only consider optimal encoding rules for the cases $m = 0$ and $m = 1$, and the efficient coding problem stated in definition (2.85), in order to verify that the conditions are both satisfied. The efficient coding problem for the case $m = 1$ is treated in Appendix 2.6, where we show that $P^*(1) = P^{DbS}_{error} < 1/2$. Using the calculations explained in Appendix 2.2, we can provide an analytical approximation to this quantity in the limiting case of large n.

Equation 2.37 states that for any encoding rule $\theta(v)$ and any prior distribution $f(v)$, the value of P_{error} for any large enough value of n will approximately equal

$$P_n(\theta; f) \;\approx\; \frac{2}{\sqrt{n\pi}} \int \hat{f}(\tilde{\theta})^2 \sqrt{\tilde{\theta}(1-\tilde{\theta})}\, d\tilde{\theta}, \qquad \text{(2.37 revisited)}$$

where $\hat{f}(\theta)$ is the probability density function of the distribution of values for $\theta(v)$ implied by the function $\theta(v)$ and the distribution $f(v)$ of values for v. In the case of DbS, the probability distribution over alternative internal representations k_i (and hence the probability of error) is the same as in the case of an encoding rule $\theta(v) = F(v)$, so that equation 2.37 can be applied. Furthermore, for any prior distribution $f(v)$, the probability distribution

of values for the quantile $z = F(v)$ will be a uniform distribution over the interval $[0, 1]$, so that $\hat{f}(\theta) = 1$ for all θ. It follows that

$$P_{\text{error}}^{\text{DbS,lim}} \approx \frac{2}{\sqrt{n\pi}} \int \sqrt{\tilde{\theta}(1-\tilde{\theta})}\, d\tilde{\theta} \;=\; \frac{1}{4}\sqrt{\frac{\pi}{n}}. \tag{2.86}$$

In the case that $m = 0$, instead, the same function $\theta(v)$ must be used regardless of the contextual distribution $f(v)$. Under the assumption that $\theta(v)$ is piecewise continuous, there must exist a magnitude v^* such that $\theta(v)$ is continuous over some interval (v_{min}, v_{max}) containing v^* in its interior. Let $\theta_{min}, \theta_{max}$ be the greatest lower bound and least upper bound respectively, such that

$$\theta_{min} \;\leq\; \theta(v) \;\leq\; \theta_{max}$$

for all $v_{min} < v < v_{max}$. The continuity of $\theta(v)$ on this interval means that by choosing both v_{min} and v_{max} close enough to v^*, we can make both θ_{min} and θ_{max} arbitrarily close to $\theta(v^*)$.

By the same argument as in Appendix 2.6, for any prior distribution $f(v)$ with support entirely contained in the interval (v_{min}, v_{max}), the pair of stimulus magnitudes v_1, v_2 will have to imply $\theta_{min} \leq \theta(v_1), \theta(v_2) \leq \theta_{max}$ with probability 1, and as a consequence the error probability $P_n(\theta; f)$ will necessarily be greater than or equal to the lower bound $P_{min}(\theta_{min}, \theta_{max})$. By choosing both v_{min} and v_{max} close enough to v^*, we can make this lower bound arbitrarily close to $P_{min}(\theta(v^*), \theta(v^*)) = 1/2$. Hence for any encoding rule $\theta(v)$ with $m = 0$, the upper bound $\bar{P}_{error}(\theta)$ cannot be lower than $1/2$. It follows that $P^*(0) = 1/2$.

Given this, condition (a) can alternatively be expressed as

$$P_{error}^{DbS} \;+\; K(1) \;<\; 1/2.$$

Note that if $K(1)$ remains less than $1/2$ no matter how large n is, this condition will necessarily be satisfied for all large enough values of n, since (2.86) implies that P_{error}^{DbS} eventually becomes arbitrarily small, in the case of large enough n. (On the other hand, the condition can easily be satisfied for some range of smaller values of n, even if $K(1) > 1/2$ once n becomes very large.)

In order to consider the conditions under which condition (b) will also be satisfied, it is necessary to further analyze the efficient coding problem stated in (2.85). We first observe that for any prior $f(v) \in \mathcal{F}$ and encoding rule

$\theta(v)$, the encoding rule can always be expressed in the form $\theta(v) = \varphi(F(v))$, where $\varphi(z)$ is a piecewise-continuous, weakly increasing function giving the probability of a "high" reading as a function of the quantile z of the stimulus magnitude in the prior distribution. We then note that when this representation is used for the encoding function in problem 2.85, the error probability $P_n(\theta; f)$ depends only on the function $\varphi(z)$, in a way that is independent of the prior $f(v)$. Hence the inner minimization problem in Eq. 2.85 can equivalently be written as

$$\inf_{\varphi(z)} P_n(\varphi). \tag{2.87}$$

This problem has a solution for the optimal $\varphi(z)$ for any number of processing units n, and an associated value, that is independent of the prior $f(v)$. Hence we can write the bound defined in (2.85) more simply as

$$\underline{P}_n = \inf_{\varphi(z)} P_n(\varphi). \tag{2.88}$$

Condition (b) will be satisfied as long as the bound defined in (2.88) is not too much lower than $P_{\text{error}}^{\text{DbS}}$. In fact, this bound can be a relatively large fraction of $P_{\text{error}}^{\text{DbS}}$. We consider the problem of the optimal choice of an encoding function $\theta(v)$ for a known prior $f(v)$ in Appendix 2.2 In the limiting case of a sufficiently large n, substitution of equation 2.2 into 2.37 yields the approximate solution

$$\underline{P}_n^{\text{lim}} \approx \frac{2}{\sqrt{n\pi}} \frac{1}{\pi^2} \frac{d\tilde{\theta}}{\sqrt{\tilde{\theta}(1-\tilde{\theta})}} = \frac{2}{\sqrt{n\pi^3}}. \tag{2.89}$$

Thus as n is made large, the ratio $\underline{P}_n^{\text{lim}} / P_{\text{error}}^{\text{DbS,lim}}$ converges to the value

$$\underline{P}^{\text{lim}} / P_{\text{error}}^{\text{DbS,lim}} = 8/\pi^2 = 0.81. \tag{2.90}$$

This means that increases in the sample size m above 1 cannot reduce $P^*(m)$ by even 20 percent relative to $P^*(1)$, no matter how large the sample may be, whereas $P^*(1)$ may be only a small fraction of $P^*(0)$ (as is necessarily the case when n is large). This makes it quite possible for $K(2) - K(1)$ to be larger than $P_{\text{error}}^{\text{DbS}} - \underline{P}$ while at the same time $P^*(0) - P_{\text{error}}^{\text{DbS}}$ is larger than $K(1)$. In this case, the optimal sample size will be $m = 1$, and the optimal encoding rule will be DbS.

While these analytical results for the asymptotic (large-n) case are useful, we can also numerically estimate the size of the terms $P^*(0), \underline{P}$, and $P_{\text{error}}^{\text{DbS}}$ in

the case of any finite value for n. We have derived an exact analytical value for $P^*(0) = 1/2$ above. The quantity $P_{\text{error}}^{\text{DbS}}$ can be computed through Monte Carlo simulation for any value of n. (Note that this calculation depends only on n, and is independent of the contextual distribution $f(v)$; we need only to calculate $P_n(\varphi)$ for the function $\varphi(z) = z$.) The calculation of $\underline{P}_n$ for a given finite value of n is instead more complex, since it requires us to optimize $P_n(\varphi)$ over the entire class of possible functions $\varphi(z)$.

Our approach is to estimate the minimum achievable value of $P_n(\varphi)$ by finding the minimum achievable value over a flexible parametric family of possible functions $\varphi(z)$. We specify the function φ in terms of the implied $\hat{F}(\theta)$, the CDF for values of $\theta(v)$. We let $\hat{F}(\theta)$ be implicitly defined by

$$[\sin\big((\pi/2)\hat{F}(\theta)\big)]^2 = g(\theta), \tag{2.91}$$

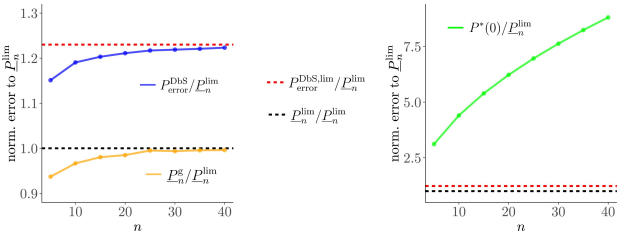
where $g(\theta)$ is a function of θ with the properties that $g(0) = 0$, $g(1) = 1$, as required for $\hat{F}(\theta)$ to be the CDF of a probability distribution. More specifically, we assume that $g(\theta)$ is a finite-order polynomial function consistent with these properties, which require that it can be written in the form

$$g(\theta) = \theta\left[1 + (\theta - 1)\left(g_0 + g_1\theta + \ldots + g_p\theta^p\right)\right], \tag{2.92}$$

where $\{g_0, \ldots, g_p\}$ are a set of parameters over which we optimize. Note that for a large enough value of p, any smooth function can be well approximated by a member of this family. At the same time, our choice of a parametric family of functions has the virtue that the CDF that corresponds to the optimal coding rule in the large-n limit belongs to this family (regardless of the value of p), since this coding rule (equation 2.3) corresponds to the case $g_0 = \ldots = g_p = 0$ of equation 2.92.

We computed via numerical simulations the best encoder function assuming $g(\theta)$ to be of order 5 (Eq. 2.92) for various finite values of $n = [5, 10, 15, 20, 25, 30, 35, 40]$, and we define the expected error of this optimal encoder for a given n to be $\underline{P}_n^g$ (i.e., a lower bound for P_n within the family of functions defined by g). Our goal is to compare this quantity to the asymptotic approximation $\underline{P}_n^{\text{lim}}$, in order to evaluate how accurate the asymptotic approximation is.

Additionally, we also compute the value $P_{\text{error}}^{\text{DbS}}$ for each finite value of n through Monte Carlo simulation (please note that $P_{\text{error}}^{\text{DbS}}$ is different from the quantity $P_{\text{error}}^{\text{DbS,lim}}$ defined in Eq. 2.86, that is only an asymptotic approximation for large n). Then, we can compare $P_{\text{error}}^{\text{DbS}}$ to the value predicted by the asymptotic approximations $P_{\text{error}}^{\text{DbS,lim}}$ and $\underline{P}_n^{\text{lim}}$.



Appendix 2.7 - Figure 2.1. Performance of efficient coding rules.

Another quantity that is important to compute, in order to determine whether DbS can be optimal when n is not too large, is the size of $P^*(0)$ relative to the quantities computed above. Since $P^*(0)$ does not shrink as n increases, it is obvious that $P^*(0)$ is much larger than the other quantities in the large-n limit. But how much bigger is it when n is small? To investigate this, we compute the value of the ratio $P^*(0)/\underline{P}_n^{\lim}$ when n is small. This quantity is given by

$$\frac{P^*(0)}{\underline{P}_n^{\lim}} = \frac{\sqrt{n\pi^3}}{4} \tag{2.93}$$

In Appendix 2.7-Figure 2.1, all error quantities discussed above are normalized relative to $\underline{P}_n^{\lim}$. The black dashed lines in both panels represent $(\underline{P}_n^{\lim}/\underline{P}_n^{\lim}) = 1$. The ratio of the asymptotic approximation for $P_{\text{error}}^{\text{DbS,lim}}$ relative to $\underline{P}_n^{\lim}$ is plotted with the red dashed lines, where $(P_{\text{error}}^{\text{DbS,lim}}/\underline{P}_n^{\lim}) \approx 1.23$. Note that the sufficient conditions for DbS to be optimal can be stated as

(a) $K(1) < P^*(0) - P_{\text{error}}^{\text{DbS}}$, and

(b) $K(2) - K(1) > P_{\text{error}}^{\text{DbS}} - \underline{P}_n$.

Therefore, Appendix 2.7-Figure 2.1 shows the numerical magnitudes of the expressions on the right-hand side of both inequalities (normalized by the value of $\underline{P}_n^{lim}$). The most important result from the analyses presented in this figure is that even for small values of n, the right-hand side of the first inequality (see right panel) will be a much larger quantity than the right-hand side of the second inequality (see left panel). Thus it can easily be the case that $K(1)$ and $K(2)$ are such that both inequalities are satisfied: it is worth increasing m from 0 to 1, but not worth increasing m to any value higher than 1. In this case, the optimal sample size will be $m = 1$, and the optimal encoding rule will be DbS.

Additionally, we found that the computations of $P_{\text{error}}^{\text{DbS}}$ for each finite value of n are slightly higher than $\underline{P}_n^{\text{lim}}$ even for small n values (blue line in the left panel), but quickly reach the asymptotic value $P_{\text{error}}^{\text{DbS,lim}}/\underline{P}_n^{\text{lim}}$ as n increases. Thus, even for small values of n, the asymptotic approximation of optimal performance for the case of complete prior knowledge is superior than DbS. We also found that the computations of $\underline{P}_n^{\text{g}}$ for each finite value of n cannot reduce $\underline{P}_n^{\text{lim}}$ by even 5 percent for small n values (orange line in the left panel). Moreover, $\underline{P}_n^{\text{g}}$ quickly reached the asymptotic value $\underline{P}_n^{\text{lim}}$, thus suggesting that the asymptotic solution is virtually indistinguishable from the optimal solution (at least based on the flexible family of g functions) also for finite values of n, which crucially are in the range of the values found to explain the data in the numerosity discrimination experiment of our study. Thus, these results confirm that the asymptotic approximations used in our study are not likely to influence the conclusions of the experimental data in our work.

Appendix 2.8: Relation to Bhui and Gershman, 2018

Bhui and Gershman [119] also argue that an efficient coding scheme can be implemented by a version of DbS. However, both the efficient coding problem that they consider, and the version of DbS that they consider, are different than in our analysis, so that our results are not implied by theirs.

Like us, Bhui and Gershman consider encoding schemes in which the internal representation r must take one of a finite number of values. However, their efficient coding problem considers the class of all encoding rules that assign one or another of N possible values of r to a given stimulus v. In their discussion of the ideal efficient coding benchmark, they do not require r to be the ensemble of output states of a set of n neurons, each of which must use the same rule as the other units, and therefore consider a more flexible family of possible encoding rules, as we explain in more detail below.

The encoding rule that solves our efficient coding problem is stochastic; even under the assumption that the prior $f(v)$ is known with perfect precision (the case of unbounded m in the more general specification of our framework, so that sampling error in estimation of this distribution from prior experience is not an issue), we show that it is optimal for the probabilities $p(k|v)$ not to all equal either zero or one. The optimal rule within the more flexible class considered by Bhui and Gershman is instead deterministic: each stimulus magnitude v is assigned to exactly one category k with certainty. The boundaries between the set of $n+1$ categories furthermore correspond to the quantiles $(1/(n+1), 2/(n+1), \ldots, n/(n+1))$ of the prior distribution, so that each category is used with equal frequency. Thus the optimal encoding rule is given by a deterministic function $y(v)$, a non-decreasing step function that takes $n+1$ discrete values.

Bhui and Gershman show that when there is no bound on m, the number of samples from prior experience that can be used to estimate the contextual distribution — their optimal encoding rule for a given number of categories N — can be implemented by a form of DbS. However, the DbS algorithm that they describe is different than in our discussion. Bhui and Gershman propose to implement the deterministic classification $y(v)$ by computing the fraction of the sampled values $\tilde{v}$ that are less than v. In the limiting case of an infinite sample from the prior distribution, this fraction is equal to $F(v)$ with probability one, and $y(v)$ is then determined by which of the intervals $[0, 1/N), [1/N, 2/N), \ldots, [(N-1)/N, 1]$ the quantile $F(v)$ falls within. Thus whereas in our discussion, DbS is an algorithm that allows each of our units to compute its state using only a single sampled value $\tilde{v}_j$,

the DbS algorithm proposed by Bhui and Gershman to implement efficient coding is one in which a large number of sampled values are used to jointly compute the output states of all of the units in a coordinated way.

Bhui and Gershman also consider the case in which only a finite number of samples $(\tilde{v}_1, \ldots, \tilde{v}_m)$ can be used to compute the representation k_i of a given stimulus magnitude v_i, and ask what kind of rule is efficient in that case. They show that in this case a variant of DbS with kernel-smoothing is superior to the version based on the empirical quantile of v_i (which now involves sampling error). In this more general case, the variant DbS algorithms considered by Bhui and Gershman make the representation k_i of a given stimulus probabilistic; but the class of probabilistic algorithms that they consider remains different from the one that we discuss. In particular, they continue to consider algorithms in which the category k_i can be an arbitrary function of v_i and a single set of m sampled values that is used to compute the complete representation; they do not impose the restriction that k_i be the number of units giving a "high" reading when the output state of each of n individual processing units is computed independently using the same rule (but an independent sample of values from prior experience in the case of each unit).

The kernel-smoothing algorithms that they consider are based on a finite set of m pairwise comparisons between the stimulus magnitude v_i and particular sampled values $\tilde{v}_j$, the outcomes of which are then aggregated to obtain the internal representation k_i. However, they allow the quantity $K(v_i - \tilde{v}_j)$ computed by comparing v_i to an individual sampled value to vary continuously between 0 and 1, rather than having to equal either 0 or 1, as in our case (where the state of an individual unit must be either "high" or "low"). The quantities $K(v_i - \tilde{v}_j)$ are able to be summed with perfect precision, before the resulting sum is then discretized to produce a final representation that takes one of only N possible values. Thus an assumption that only finite-precision calculations are possible is made only at the stage where the final output of the joint computation of the processors must be “read out”; the results of the individual binary comparisons are assumed to be integrated with infinite precision. In this respect, the algorithms considered by Bhui and Gershman are not required to economize on processing resources in the same sense as the class that we consider; the efficient coding problem for which they present results is correspondingly different from the problem that we discuss for the case in which m is finite.

3

EFFICIENT NUMEROSITY ESTIMATION UNDER LIMITED TIME

J. A. Heng, M. Woodford, R. Polanía. Efficient numerosity estimation under limited time. In review.

Contributions

Conceptualization, Formal analysis, Data analysis, Investigation, Visualization, Writing.

3.1. Abstract

The ability to rapidly estimate non-symbolic numerical quantities is a well-conserved sense across species with clear evolutionary advantages. Despite its importance, the rapid representation and estimation of numerosity is surprisingly imprecise and biased. However, a formal explanation for this seemingly irrational behavior remains unclear. We develop a unified normative theory of numerosity estimation that parsimoniously incorporates in a single framework information processing constraints alongside Brownian diffusion noise to capture the effects of exposure time of sensory estimations, logarithmic encoding of numerosity representations, and optimal inference via Bayesian decoding. We show that for a given allowable biological capacity constraint our model naturally endogenizes time perception during noisy efficient encoding to predict the complete posterior distribution of numerosity estimates. This model accurately predicts many features of human numerosity estimation as a function of temporal exposure, indicating that humans can rapidly and efficiently sample numerosity information over time. Additionally, we demonstrate how our model fundamentally differs from a thermodynamically-inspired formalization of bounded rationality, where information processing is modeled as acting to shift away from default states. The mechanism we propose is the likely origin of a variety of numerical cognition patterns observed in humans and other animals.

3.2. Introduction

The ability to rapidly represent and estimate non-symbolic numerical quantities is a fundamental cognitive function for behavior in humans and other animals, which may have emerged during evolution to support fitness maximization [48]. Since the properties of numerosity estimation started to be studied nearly a century ago, it has been commonly observed that the representation and estimation of numerical quantities are imprecise and biased [136]. Despite the importance of numerosity estimation for various cognitive processes and ultimately survival, the questions remain: what are the origins of the observed variability and biases in numerosity estimations? Are these deviations efficient and predictable when organisms are urged to rapidly estimate numerical quantities?

Extensive empirical research in the representation and estimation of non-symbolic numerical quantities has consistently reported and studied various features that characteristically emerge during numerosity estimation, including: (i) subitizing small numbers [137]; (ii) overestimation of small numbers (outside the subitization range) and underestimation of large numbers [138], with especially biased estimates in the case of larger numbers [133]; (iii) a coefficient of variation that is approximately constant across all numerosities, a property termed scalar variability [139]; and (iv) estimation acuity modulated by duration of stimulus presentation and sensory reliability [140]. But do all the above-mentioned behavioral patterns have a common origin?

Organisms do not have unlimited biological resources or unlimited time to process sensory information from the environment, and neural computations are metabolically expensive [141]. Thus, it has been suggested that the observed variability and biases in our estimations of our sensory world emerge from fundamental principles of acquiring information from environmental regularities that should ultimately lead to developing efficient behavioral strategies [37, 72, 73, 86, 142]. Here we argue that all the above-mentioned behavioral features emerging during numerosity estimation have a common origin: given biological constraints on information acquisition, numerosity estimation emerges from a system that efficiently considers, first, prior knowledge of the environment, second, information of the current numerosity being evaluated, and third, the amount of time (or sensory reliability) available to process such information.

We develop a unified normative model of numerosity estimation that parsimoniously incorporates information constraints together with long modeling

traditions of human and animal psychophysical performance in psychology and neuroscience: (i) Brownian diffusion noise to capture the effects of time exposure of sensory information [143], (ii) logarithmic encoding of numerosity representations [83], and (iii) optimal Bayesian decoding. As a result, we show that for a given allowable biological capacity constraint, our model naturally incorporates time (or sensory reliability) perception during noisy efficient encoding to predict the corresponding posterior distribution of numerosity estimates via optimal Bayesian decoding. Here we refer to our approach as the "sequential-encoding/Bayesian-decoding" model, henceforth SEB.

We also consider a second well-known approach for studying bounded rationality inspired by principles of thermodynamics and statistical physics. This family of models assumes that given a default state (e.g., a default distribution over possible responses) and a sensory stimulus, the observer acts in a way such that they attempt to shift from the default state to a new state that matches as closely as possible the value of the sensory stimulus. Bounded rationality comes into play in the case of acting when only a given amount of change in information (energy invested) between the default and new state can be afforded. This class of models has been used in a wide range of applications [144–147], including recently to study how perceptual estimation under limited time relates to cognitive capacity and action responses [50]. Here we refer to this class of models as the "thermodynamically inspired model", henceforth TIM.

Here, we will formally demonstrate that the two approaches that we consider here (SEB and TIM) are in fact classes of models with completely different views on bounded rationality, which run the risk to be confused. On the one hand, variability in the estimation responses in SEB is attributed to *sensing* costs, which generate noisy sensory encoding. On the other hand, in instantiations of TIM applied to sensory estimation, variability is generated by *acting* costs during response selection. Crucially, here we demonstrate that these two approaches applied to numerosity estimation lead to apparently similar but distinguishable quantitative and qualitative predictions that are identifiable and falsifiable. Our empirical tests applied to a large numerosity estimation data set provide a clear indication humans can rapidly and efficiently sample numerosity information over time via an efficient noisy encoding and decoding process.

3.3. Results

The presentation of our results is divided into three parts: First, we present our sequential-encoding/Bayesian-decoding model (SEB) which parsimoniously endogenizes perceptual exposure times in its likelihood function alongside parameters of the prior distribution for a given biological capacity bound. Second, we introduce the thermodynamically inspired model (TIM) applied to sensory estimation, and compare it with the SEB model. Third, we apply rigorous quantitative and qualitative model evaluations based on a large publicly available human numerosity estimation dataset (n=400 participants across four different experiments).

A Bayesian model of numerosity estimation

Extensive behavioral and physiological work studying the representation of both non-symbolic and symbolic numerical quantities strongly suggests that internal representations r can be assumed to be encoded by a quantity that is proportional to the logarithm of the number n plus stimulus-independent random error [83, 85, 133]. However, a key contribution of our work is to formally study how these perceptual errors may depend on stimulus duration t of the form

$$r \sim N(\log n, \nu^2(t)). \tag{3.1}$$

We assume the prior distribution to be a log-normal distribution from which the true numerosity n is drawn to be

$$p(\log n) \sim N(\mu, \sigma^2). \tag{3.2}$$

While the distribution of various quantities in linguistics, economics, and ecology appears to be well-described by log-normal distributions [148], others have argued that power-law distributions approximately describe the empirical frequency of numbers in natural environments [149, 150]. We note, however, that the two-parameter family of possible log-normal prior distributions includes as a limiting case the power-law distributions (Supplementary Note 3.1). If we consider a normalized prior of the form

$$p(n) \propto \exp\big(-\alpha(\log n) - \gamma(\log n)^2\big), \tag{3.3}$$

for some parameters α, γ with $\gamma \geq 0$. If $\gamma > 0$, this corresponds to a log-normal prior, with $\mu = (1-\alpha)/(2\gamma)$, $\sigma^2 = 1/(2\gamma)$. If instead $\gamma = 0$ but $\alpha > 0$, this corresponds to a power-law prior

$$p(n) \propto n^{-\alpha}. \tag{3.4}$$

Thus, our model allows for the possibility that encoding and decoding are adapted to different priors that are learned for different contexts, rather than a single process being used in all contexts.

A log-normal (or power-function) prior assumption implies that the posterior distribution for n conditional on the noisy measurement r will be also log-normal (Supplementary Note 3.1). Here we assume that the numerosity estimate $\hat{n}$ minimizes the MSE when stimuli are drawn from the prior distribution. This implies that conditional on n, the estimate $\hat{n}$ will be log-normally distributed (Supplementary Note 3.1)

$$p(\log \hat{n} \mid n) \sim N(\hat{\mu}(n,t), \hat{\sigma}^2(t)), \tag{3.5}$$

where $\hat{\mu}(n,t)$ is an affine function of $\log n$, and $\hat{\sigma}^2(t)$ is independent of n. However, both $\hat{\mu}$ and $\hat{\sigma}^2$ may depend on temporal numerosity processing t, as we formally elaborate below.

Exposure time and precision of internal representations

We suppose now that the internal representation r consists of the sample path of a Brownian motion z_s over a time interval $0 \leq s \leq \tau$, starting from an initial value $z_0 = 0$. The drift m of the Brownian motion is assumed to depend on n, while its instantaneous variance ω^2 is independent of n; the length of time τ for which the Brownian motion evolves is also independent of n, but depends on the viewing time t. In assuming sensory evidence given by a Brownian motion with a drift that depends on the stimulus, we follow a long modeling tradition that includes the popular drift-diffusion model [143]. Models of this kind have been used since the late 60s to account quantitatively for the way in which the accuracy of perceptual judgments is affected by manipulations of viewing time [151].

More specifically, we assume that m is an affine transformation of the logarithm of n,

$$m \;=\; \xi + \psi \log n, \tag{3.6}$$

where the parameters ξ and ψ may depend of the statistics of a particular environment. We suppose that the choice of these coefficients is subject to a "power constraint" which requires the average value of m^2 to be within some finite bound

$$\mathrm{E}[m^2] \;\leq\; \Omega^2 \;<\; \infty. \tag{3.7}$$

This bound on the amount of variation in the drift limits the precision with which different stimuli can be perceived for any given τ. The value of τ is

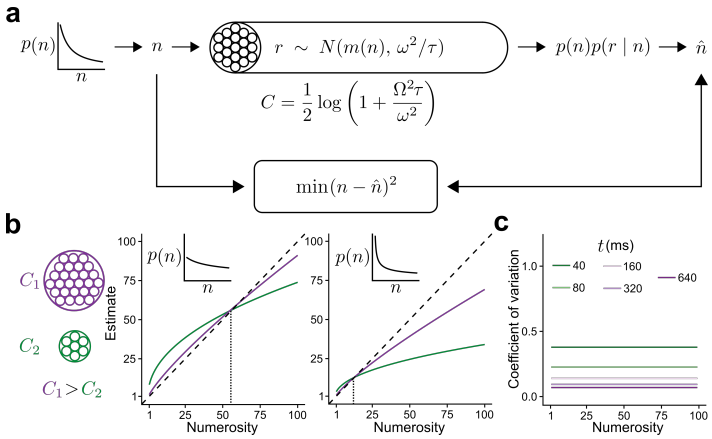


Figure 3.1: **Overview of the SEB model. a)** Schematic description of the SEB model. A numerosity n is drawn from a stationary environment known by the observer. The observer has a limited capacity C to represent the numerosity. The internal representation r is a random draw from a Gaussian distribution, the mean of which depends on n but the variance does not. The observer then infers the estimate $\hat{n}$ based on the representation r and the prior distribution of n as to minimize the MSE between the estimate and the numerosity. **b)** Illustration of the predictions for an observer with a high (purple) or low (green) channel capacity where n is drawn from a distribution with a high (left) or low (right) variance. All curves exhibit overestimation for lower numerosities and underestimation for higher numerosities. However, these biases are reduced in the case of high capacity. The crossover point between under- and overestimation increases with the variance of the numerosity distribution. **c)** Illustration of the coefficient of variation (i.e., $\mathrm{SD}[\hat{n}]/\mathrm{E}[\hat{n}]$) for different capacities. The coefficient of variation is independent of the numerosity and decreases with capacity, which is dependent on the viewing time t.

assumed to grow linearly with the viewing time, up to some finite bound B,

$$\tau \;=\; \min(t,\, B), \tag{3.8}$$

representing a constraint on the amount of evidence that can be maintained in working memory. The latter bound constrains the degree to which precision can be increased by further increases in viewing time.

Under the assumption that the particle position under Brownian motion is normally distributed with its parameters evolving as a function of τ, one can show that r is a draw from the distribution (Supplementary Note 3.2)

$$r \;\sim\; N(m(n),\, \omega^2/\tau). \tag{3.9}$$

This effectively states that r can be seen as the output of a *Gaussian channel* with input m [152]; hence the problem of optimally choosing the function $m(n)$ is equivalent to an optimal encoding problem for a Gaussian channel. The capacity C of such a channel is a quantitative upper bound on the amount of information that can be transmitted regardless of the encoding rule, which is equal to (Supplementary Note 3.3)

$$C = \frac{1}{2}\log\left(1 + \frac{\Omega^2\tau}{\omega^2}\right). \tag{3.10}$$

Note that in our model the channel capacity C grows as a logarithmic function of τ because the correlation of successive increments in the encoding by a Brownian motion prevents the information content from growing in proportion to such increments.

Here we assume that the goal is to design a capacity-limited system that minimizes the mean squared error (MSE) of the estimate $\hat{n}$ when n is drawn from a log-normal prior distribution. It is possible to show that in our optimization problem, which assumes a channel with "power transmission" constraint Ω^2, the encoding noise ν in Eq. 3.1 is given by (see Supplementary Note 3.2 for proof)

$$\nu(t) = \frac{1}{\sqrt{\tau}}\frac{\omega}{\Omega}\cdot\sigma. \tag{3.11}$$

That is, encoding precision grows with viewing time t. Recall that σ is the variance of the log-normal prior, and therefore the solution reveals that the likelihood is independent of parameter μ of the log-normal prior distribution, but depends on the second moment of this prior distribution and viewing

time t. Defining $R \equiv \Omega/\omega$, the noise of numerosity encoding is given by $\nu(t) = 1/G$, where $G = \min(R\sqrt{t}/\sigma, B)$ and B a maximum biologically allowed bound on sensory precision.

These results lead to the following predictions from our model: (i) $\mathrm{E}[\hat{n} \mid n]$ is a concave function of n with overestimation for small numbers (when these are not so small that the discreteness of available responses leads to nearly-deterministic responses), but underestimation for large numbers (Fig. 3.1b and Supplementary Note 3.2). (ii) The crossover point from overestimation to underestimation changes as a function of the numerosity range, and in each context, the concavity of $\mathrm{E}[\hat{n} \mid n]$ depends on the amount of resources available to perform the numerosity estimation task. This prediction was clearly confirmed in a previous empirical work [138]. (iii) Because of the discreteness in the set of responses, there is predicted to be little variability in responses in the case of low enough numbers, and the subitizing-like range for small numbers becomes larger as the biological capacity C or the viewing time t increases. (iv) For numbers beyond the subitizing-like range, based on the properties of the log-normal distribution, it can be shown that the coefficient of variation (Supplementary Note 3.1)

$$\frac{\mathrm{SD}[\hat{n}]}{\mathrm{E}[\hat{n}]} = \sqrt{e^{\hat{\sigma}^2(t)} - 1} \tag{3.12}$$

does not depend on the input numerosity n, thus delivering the property of scalar variability, irrespective of n [133], but here we show that this coefficient will depend on time exposure t, with the predicted constant coefficient of variation decreasing as t gets larger proportionally with $\sqrt{e^{\hat{\sigma}^2(t)} - 1}$ (Fig. 3.1c).

A thermodynamically inspired model of bounded rationality

Here we briefly introduce a popular approach to studying systems with bounded capacity across domains in human cognition and machine learning: a thermodynamically inspired formalization where information processing is modeled as changes from a default state, which come at some energetic cost, that can be quantified by differences in free energy. This class of models can be applied for the case where an observer intends to minimize some form of expected loss (the case we study here for the case of estimation error minimization), subject to information constraints [144]. More formally, let $q(\hat{n})$ be a default state (distribution) over possible responses $\hat{n}$ in a given environment or context. When presented with a stimulus n, the resource-constrained observer attempts to transform the initial state q into a new

state of possible responses $p(\hat{n} \mid n)$. This transformation of states can be modeled as the optimization of the free energy functional

$$F[p(\hat{n} \mid n)] := - \underbrace{\mathbb{E}[L; \hat{n}]}_{\text{Expect. Loss}} \underbrace{- \frac{1}{\beta} D_{KL}\left(p(\hat{n} \mid n) \| q(\hat{n})\right)}_{\text{Constrained State Change}} \tag{3.13}$$

where L is a loss function, for instance, the squared error $(\hat{n} - n)^2$. The second term is the Kullback-Leibler divergence between q and $p(\hat{n} \mid n)$, where β trades off the relative importance of changing from the default state q, thus determining the resources that the observed invests in the estimation task. The goal is to find the optimal distribution of responses

$$p^*(\hat{n} \mid n) := \arg\max_{p(\hat{n}|n)} F[p(\hat{n} \mid n)]. \tag{3.14}$$

The optimal distribution of responses in this variational problem has an analytical solution of the form

$$p^*(\hat{n} \mid n) \propto q(\hat{n}) \exp\left(-\beta h(L_n(\hat{n}))\right), \tag{3.15}$$

where h is a function of L and potentially other elements incorporated in the expected loss function in Eq. 3.13.

TIM applied to numerosity estimation

A recent work applied a model from the TIM family to study a resource-constrained model of human numerosity estimation [50]. This is also a formulation of how the distribution of reported numerosity estimates $\hat{n}$ of a stimulus magnitude should vary depending on the true stimulus n. This can be stated generally as the hypothesis that conditional on n the response distribution $p(\hat{n} \mid n)$ is the probability distribution over a set of possible responses N that minimizes the mean squared error (MSE), subject to the constraint

$$D_{KL}\left(p_n \| q\right) \leq C(t) = \min(Rt, B), \tag{3.16}$$

where $C(t)$ is a positive bound that depends on the amount of time t for which the stimulus is presented. This formulation can be interpreted as a model in which errors in the observer's responses can be attributed to a "cost of control" of the responses: it is difficult for the observer to give responses different from the default state q, though their response distribution to the

individual stimulus n is optimal given a constraint on the possible precision of their responses.

Similar to our SEB model, in TIM it is assumed that perception extracts information linearly in time at a rate R until an overall capacity bound B is reached. The goal is to find the distribution of numerosity estimates $p^*(\hat{n} \mid n)$ that minimizes the mean squared error

$$\mathrm{MSE} \equiv \sum_n q(n) \sum_{\hat{n}} p(\hat{n} \mid n)(\hat{n} - n)^2 \tag{3.17}$$

under the constraint given in Eq. 3.16.

The optimization problem described above yields the following analytical solution [50]

$$p^*(\hat{n} \mid n) \propto q(\hat{n}) \exp\left(-\beta_n q(n)(n - \hat{n})^2\right), \tag{3.18}$$

where β_n is chosen to satisfy the bound in Eq. 3.16. Note that this solution has the familiar form obtained in Eq. 3.15 with L as the loss function $L_n = (n - \hat{n})^2$.

While this solution is usually linked to a bounded-rational Bayesian computation (given the observation that the default distribution q is multiplied by a function of n given $\hat{n}$), here we clarify that this solution does not correspond to a Bayesian inference process with noisy sensory percepts. In fact, the TIM formulation assumes that the perception of the sensory stimulus n is noiseless, and all the variability observed during the estimation process is related to the cost of acting accurately, that is, a cost in the precision of response selection when shifting away from the default state. Note that this is fundamentally different from the SEB model, in which all the estimation variability is attributed to noisy sensory encoding.

General similarities and differences between SEB and TIM

We elaborated an illustrative example that allows the predictions of the two models to be solved analytically, thus allowing us to understand the key differences between the two models (Supplementary Note 3.4). These analyses reveal some similarities between the predictions of the two models, however, there are also notable differences. First, while both models predict that biases decrease in general for larger viewing times, TIM implies a faster decrease rate in such biases as t increases. Second, for a given input stimulus n, the two models do not imply that $\mathrm{var}[\hat{n} \mid n]$ co-varies with the bias in the same way. As $t \to 0$, the Bayesian model implies that the variance forecasts should

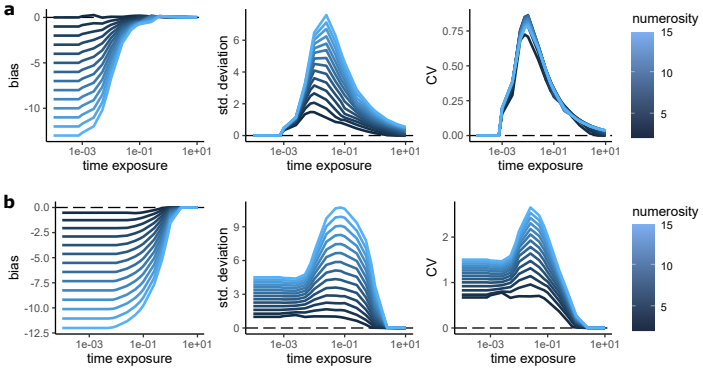


Figure 3.2: **General similarities and differences between SEB and TIM. a)** Computation of the bias ($\mathrm{E}[\hat{n} \mid n] - n$, left), standard deviation ($\mathrm{SD}[\hat{n} \mid n]$, middle), and the coefficient of variation ($\mathrm{CV} = \mathrm{SD}[\hat{n} \mid n]/\mathrm{E}[\hat{n} \mid n]$, right) as a function of different time exposures t for different numerosities n (color scale of the solid lines) in the SEB model. **b)** Same as panel a, but this time computed for TIM. Differences between the two models are particularly salient in the computation of the SD and the CV.

fall to zero; TIM implies that this is the case in which estimates should be more variable: the variance of the prior distribution.

These analytical insights were studied over all possible responses in the continuous space and do not directly apply to numerosity estimation. Therefore, we conducted numerical analyses to study whether the same signatures emerge in SEB and TIM when the solutions are restricted over the space of positive integers. As expected, both models predict that biases decrease in general for larger viewing times, and mirroring the results of the analytical solution, TIM reveals a faster decrease in the bias rate as t increases (Fig. 3.2, left panels). Moreover, as $t \to 0$, SEB implies that the variance forecasts should fall to zero, but this is not the case in TIM where the predicted variability of estimates is clearly larger (Fig. 3.2, middle panels). Finally, the computation of the coefficient of variation ($\mathrm{CV}[\hat{n}] \equiv \mathrm{SD}[\hat{n}]/\mathrm{E}[\hat{n}]$) reveals that in SEB this metric is nearly identical for all numerosities n irrespective of time exposure t, thus reflecting the scalar variability effect (Fig. 3.2a, right panel). In TIM, however, the scalar variability phenomenon is absent irrespective of time exposure t. These differences make the two models different and identifiable and generate somewhat different qualitative predictions.

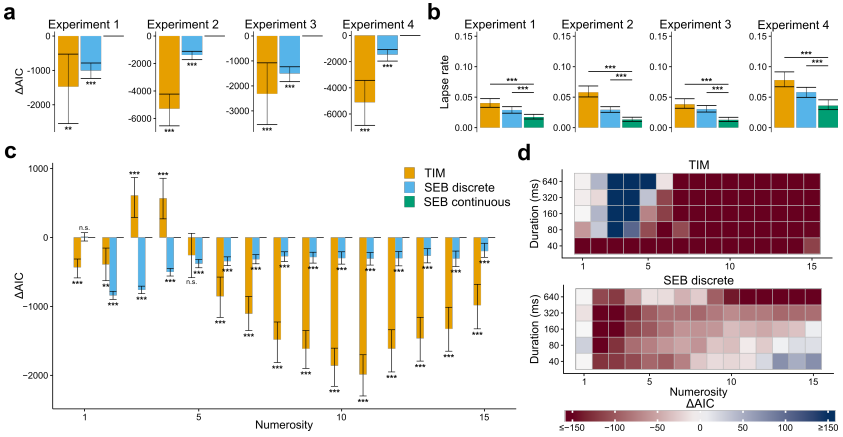


Figure 3.3: **The SEB model quantitatively outperforms the TIM model when the prior parameters are fixed. a)** Difference in AIC between the SEB continuous model (green) and the TIM model (orange) or the SEB discrete (blue). The ΔAICs were computed for each participant and summed. The error bars represent the 95% confidence interval based on bootstrapping of the participants' ΔAICs. The SEB continuous model outperforms both the TIM model (experiment 1: $(T(99) = 2.86, p < 0.01)$, experiment 2: $(T(99) = 8.70, p < 0.001)$, experiment 3: $(T(99) = 3.61, p < 0.001)$, experiment 4: $(T(99) = 5.79, p < 0.001)$) and the SEB discrete model (experiment 1: $(T(99) = 8.64, p < 0.001)$, experiment 2: $(T(99) = 9.16, p < 0.001)$, experiment 3: $(T(99) = 10.1, p < 0.001)$, experiment 4: $(T(99) = 6.69, p < 0.001)$). **b)** Average lapse rate parameter per participant for each model and experiment. The error bars represent the 95% confidence interval based on bootstrapping of the participants' lapse rate. The lapse rate of the SEB continuous model is lower than the TIM model (experiment 1: $(T(99) = 6.33, p < 0.001)$, experiment 2: $(T(99) = 9.86, p < 0.001)$, experiment 3: $(T(99) = 6.91, p < 0.001)$, experiment 4: $(T(99) = 5.75, p < 0.001)$) and the SEB discrete model (experiment 1: $(T(99) = 6.58, p < 0.001)$, experiment 2: $(T(99) = 7.11, p < 0.001)$, experiment 3: $(T(99) = 8.63, p < 0.001)$, experiment 4: $(T(99) = 6.80, p < 0.001)$). These results indicate that less variability is associated to lapses of attention in the SEB continuous model, which suggests a better fit to behavior. **c)** Difference in AIC between the SEB continuous model and the TIM model or the SEB discrete model for each numerosity. The error bars represent the 95% confidence interval based on bootstrapping of the participants' AICs. The SEB continuous model outperforms the TIM model except for numerosities 3, 4 and 5 and the SEB discrete model for all numerosities except numerosity 1. **d)** AIC differences between the SEB continuous model and the TIM model (top) and the SEB discrete model (bottom) for all experiments shown for different numerosities and levels of sensory evidence (stimulus presentation duration or contrast). Duration values are assigned to Weber contrasts of experiment 4 for pooling purposes (40ms–10%, 80ms–20%, 160ms–40%, 320ms–80%, 640ms–160%). The SEB continuous model outperforms the TIM and SEB discrete models for most numerosities and levels of sensory evidence.

SEB largely outperforms TIM applied to human numerosity estimation

We now compare TIM with SEB models using the experimental data of a pre-registered study provided in previous work [50] (see Methods). In brief, on each trial, between 1 and 15 dots were flashed, followed by a noise mask. The participants were then asked to type their estimation of how many dots were displayed. There were three between-participant experiments (n=100 per experiment) that manipulated available stimulus information (variable exposure time: $t \in [40, 80, 160, 320, 640]$ ms) and different ways of controlling non-numerical properties of the stimuli: the average dot size (experiment 1), surface density (experiment 2) or surface area (experiment 3) of the dots.

To fully constrain inference solely to the normative solutions of stimulus exposure derived above for both SEB and TIM, we fixed the prior distribution before fitting the behavioral data to a prior equivalent of the form $1/n^{\alpha}$ power-law. It has previously been argued that the prior probability of how often numerosities are encountered and represented roughly follows a $1/n^{\alpha=2}$ power-law distribution [149, 150]. Thus, a priori, we choose $\alpha = 2$, following the same assumption adopted in previous work [50]. By fixing such ecologically valid prior, we alleviate the critique of allowing an arbitrary choice of prior and likelihood functions to fit inference models to the data, as a consequence of which it is sometimes argued that their predictions are potentially vacuous [153]. Nevertheless, it is well possible that each individual has learned their own distribution during their lifespan [154, 155]. Therefore, we also considered a more flexible class of models where we allowed the parameters of the prior distribution to be free parameters alongside the capacity constraint and capacity bound.

We considered two possible ways of inferring the numerosity estimates based on the SEB approach (methods): (i) using the analytical solutions over the continuous positive real line, and (ii) using discrete encoding and decoding restricted to the positive integer numbers, thus similar in nature to the TIM specification. Finally, we considered a guessing rate g in the model fits, which assumes that on g proportion of trials, participants were distracted and had no information about the number of dots in the display, meaning that their estimate was effectively a random sample from their prior. Thus, both numerosity estimation models SEB and TIM have exactly the same degrees of freedom (the capacity constraint, capacity bound, and g), in addition to the prior parameters in the flexible class of models.

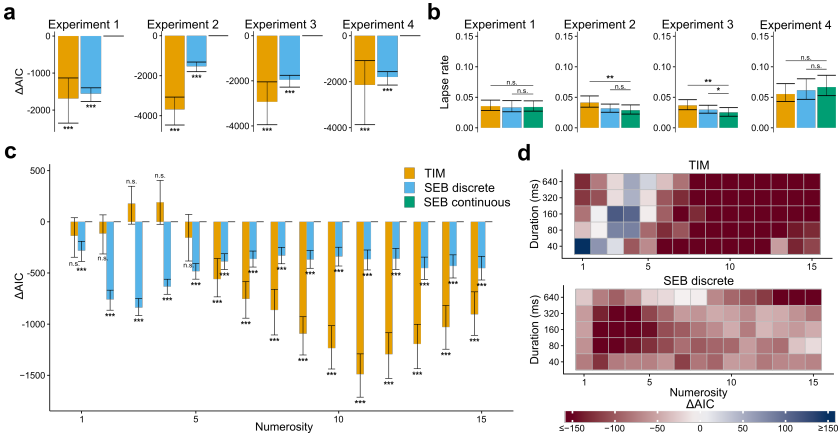


Figure 3.4: **The SEB model quantitatively outperforms the TIM model when the prior parameters are free. a)** Difference in AIC between the SEB continuous model (green) and the TIM model (orange) or the SEB discrete (blue). The ΔAICs were computed for each participant and summed. Error bars represent the 95% confidence interval based on bootstrapping of the participants' ΔAICs. The SEB continuous model outperforms both the TIM model (experiment 1: $(T(99) = 5.57, p < 0.001)$, experiment 2: $(T(99) = 10.40, p < 0.001)$, experiment 3: $(T(99) = 6.29, p < 0.001)$, experiment 4: $(T(99) = 3.54, p < 0.001$ and the SEB discrete model (experiment 1: $(T(99) = 16.04, p < 0.001)$, experiment 2: $(T(99) = 12.35, p < 0.001)$, experiment 3: $(T(99) = 15.46, p < 0.001)$, experiment 4: $(T(99) = 11.83, p < 0.001)$). **b)** Average lapse rate parameter per participant for each model and experiment. Error bars represent the 95% confidence interval based on bootstrapping of the participants' lapse rate. The lapse rate of the SEB continuous model is lower than the TIM model for experiment 2 $(T(99) = 2.95, p < 0.01)$ and experiment 3 $(T(99) = 2.95, p < 0.01)$ but not for experiment 1 $(T(99) = 0.34, p = 0.734)$ and experiment 4 $(T(99) = 1.72, p = 0.088)$ and the SEB discrete model for experiment 3 $(T(99) = 2.118, p < 0.05)$ but not for the other experiments (experiment 1: $(T(99) = 0.33, p = 0.743)$, experiment 2: $(T(99) = 1.416, p = 0.160)$, experiment 4: $(T(99) = 1.33, p = 0.187)$). **c)** Difference in AIC between the SEB continuous model and the TIM model or the SEB discrete model for each numerosity. Error bars represent the 95% confidence interval based on bootstrapping of the participants' AICs. The SEB continuous model outperforms the TIM model except for numerosities 1 to 5 and the SEB discrete model for all numerosities. **d)** AIC differences between the SEB continuous model and the TIM model (top) and the SEB discrete model (bottom) for all experiments shown for different numerosities and levels of sensory evidence (stimulus presentation duration or contrast). Duration values are assigned to Weber contrasts of experiment 4 for pooling purposes (40ms–10%, 80ms–20%, 160ms–40%, 320ms–80%, 640ms–160%). The SEB continuous model outperforms the TIM and SEB discrete models for most numerosities and levels of sensory evidence.

Quantitative model comparison

For each experiment where stimulus presentation time t was manipulated, we fit both types of model to the data of each participant (Methods). In parameter recovery exercises we found that all model parameters are identifiable and this is also confirmed by the weak relationship between parameters across participants (Supplementary Fig. 3.1). We first examined the restricted models where the prior is fixed $1/n^2$. For experiment 1, we found that the difference in Akaike information criterion (AIC) favoured SEB, where the continuous version of SEB had a clear advantage over TIM: ΔAIC=1472 [95%-CI 570-2553] in favor of SEB (paired t-test: $T(99) = 2.86, p < 0.01$. For experiment 2 (dot density controlled), the difference in AIC is 5284 [95%-CI 4185-6690] in favor of SEB ($T(99) = 8.70, p < 0.001$). For experiment 3 (dot area controlled), the difference in AIC is 2316 [95%-CI 1218-3686] in favor of SEB ($T(99) = 3.61, p < 0.001$, see Fig. 3.3a). In addition, the SEB continuous model provided better fits than its discrete version ($T(99) \geq 8.64, p < 0.001, d > 0.86$ Δ AIC$\geq$ 997).

Previous theoretical and empirical work suggests that two ways in which the amount of information available to process information can be studied are by manipulating time exposure and also by changing stimulus contrast [72]. Thus, we also considered this alternative way of manipulating sensory reliability, which should affect the channel capacity transmission (see Eq. 3.10). To test this, we analyzed data of a numerosity estimation experiment, where in each trial the visual contrast of numerosity was manipulated at a constant presentation time (n=100 participants, experiment 4, Methods). We found that also in this experiment the SEB-continuous model fits the data better than TIM ($\Delta AIC = 5106$; [95%-CI 3452-6880] ($T(99) = 5.79, p < 0.001$), Fig. 3.3a) and the discrete version of SEB ($\Delta AIC = 1453$; [95%-CI 1059-1907] ($T(99) = 6.69, p < 0.001$)).

To make sure that the overall quantitative differences were not driven by a few numerosities, we computed the difference in AIC for each numerosity and each model. We found a significant interaction models*numerosity of the ΔAICs ($F(28, 16758) = 7.84$, $p < 0.001$) with post hoc tests revealing that this effect was more pronounced for higher numerosities (SEB continuous vs TIM: paired t-tests $p < 0.001$ for numerosities $n > 5$, Fig. 3.3b) and also for $n \in [1, 2]$ (paired t-tests $p < 0.01$). The relative advantage of the TIM model for $n \in [3, 4]$) at large presentation times t might be explained by the fact that smaller numerosities are close to the subitizing range and therefore most of the posterior density mass is concentrated around the

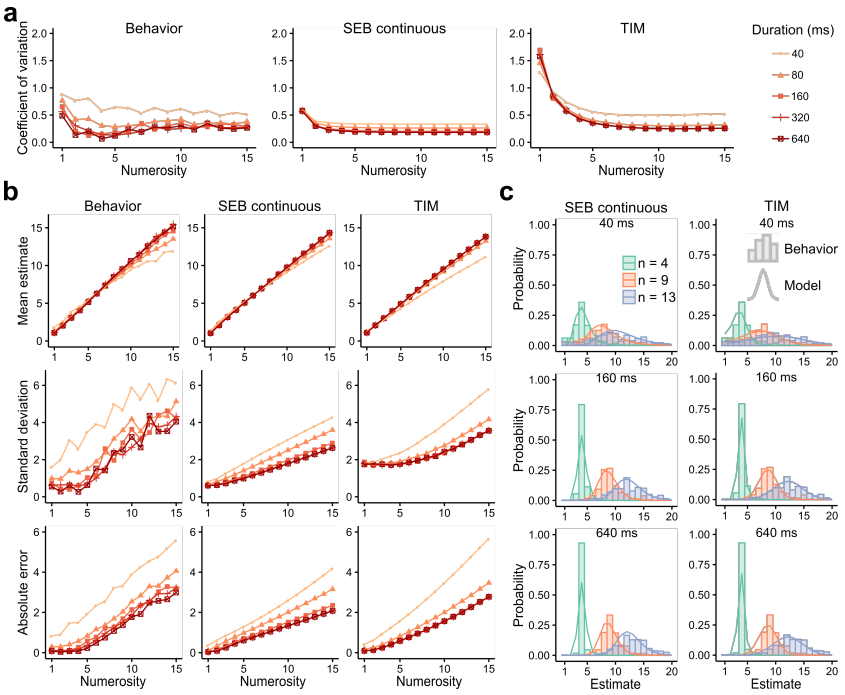


Figure 3.5: **The SEB continuous model with free prior parameters qualitatively explains behavior. a)** Coefficient of variation (SD[$\hat{n}$]/E[$\hat{n}$]) of the behavior data (left) and predicitons of the SEB model (middle) and TIM model (right) using a prior with free parameters for different numerosities and stimulus presentation duration. Predictions were performed by taking for each parameter the value with this highest density across participants. Duration values are assigned to Weber contrasts of experiment 4 for pooling purposes (40ms–10%, 80ms–20%, 160ms–40%, 320ms–80%, 640ms–160%). The TIM model predicts a higher CV for lower numerosities. This feature is not present in the behavior data nor the SEB predictions. **b)** Mean estimate (top), standard deviation (middle) and absolute error (bottom) of the behavior data (left) and predictions of the SEB model (middle) and TIM model (right). **c)** Posterior distribution of estimates to numerosities 4 (green), 9 (red) and 13 (blue) of the SEB (left) and the TIM (right) model for different stimulation presentation duration (40ms (top), 160ms (middle), 640ms (bottom)). Behavior of participants is shown as histograms. For visualization purposes, estimates above 20 are not shown.

input n, which is better explained by the TIM model as this model has a tendency to subitize more strongly at small numerosities [50]. Interestingly, for $n \in [1, 2]$, the Bayesian model predicts noisier estimations (in particular for smaller exposure times t) which are not supported by the TIM, with the AICs favoring the former.

Additionally, we inspected the AIC differences split by both numerosity and sensory evidence (time or contrast), finding a similar pattern, but the differences were larger for small levels of sensory evidence. Thus, SEB appears to be more sensitive to capturing behavior for stimuli generating higher noise levels in the encoding operations.

Moreover, we compared the guessing rates g between the two kinds of models. Guessing rates can capture unassigned variance in miss-specified models, thus we conjectured that a relatively smaller value of g would provide further evidence for better mechanistic fits captured by the best model. While the guessing rates are overall small (suggesting that the amount of distractions during task performance was minimal), we found that guessing rates were systematically smaller in the SEB model ($T(99) \geq 5.75, p < 0.001, d > 0.58$ for each experiment, Fig. 3.3b and Supplementary Table 3.1). Thus, while the effects of distraction are estimated to be relatively small in both models, our analyses provide a clear indication that potentially unassigned variance due to distraction is lower in the SEB model relative to TIM.

We repeated the same set of analyses treating parameters of the log-normal prior as free parameters. The results of these analyses mirrored the initial analyses. That is, (i) we found that the SEB model fit the data better than TIM in all four experiments ($T(99) \geq 3.54, p < 0.001$), Fig. 3.4a), (ii) the continuous version of SEB performed better in general than its discretized version (Fig. 3.4a) and (iii) the guessing rates were significantly smaller in the SEB model than the TIM model for experiments 2 and 3 ($T(99) \geq 2.95, p < 0.01$ but not for experiments 1 and 4, Fig. 3.4b).

The next question to ask is whether the models with free prior parameters outperformed the models with the prior fixed to $1/n^2$. We found that for each model considered here, the models with free prior parameters outperformed their corresponding version with fixed parameters ($T(99) \geq 5.62, p < 0.001$, Supplementary Table 3.2). Additionally, accounting for population variability in the quantitative metrics between participants across all models considering here, BMS reveals that the Bayesian model with free prior parameters is clearly favored relative to all the other models for experiments 1, 2 and 3 ($P_{\text{xp}} > 0.99$ for each experiment) but equally favored to the TIM model with

free prior parameters for experiment 4 ($P_{\text{xp}} = 0.50$). These results allow us to conclude two important points. First, variability in the prior parameters of the prior distribution is key to more accurately explaining human numerosity estimations. Second, our results provide a clear indication that the effects of temporal time exposure are better captured by the noisy encoding model (SEB) relative to an action control-like model (TIM).

Qualitative predictions

We first examined the qualitative features of scalar variability in both data and the predictions of the SEB continuous and the TIM models with free prior parameters. For each numerosity value, we computed the coefficient of variation (CV: $\text{SD}[\hat{n}]/\text{E}[\hat{n}]$). We found that the empirical data follows the previously observed properties of scalar variability for numerosities greater than 4 (i.e., a flat CV irrespective of numerosity and sensory evidence), with a slight systematic increase of CV for smaller numbers (Fig. 3.5a left). This relative CV increase for small numbers could be explained by the presence of small lapse rates g which have a greater impact on the CV for small n. We found that the SEB model accounts for these qualitative observations (Fig. 3.5a middle), however, the TIM model generates slightly different predictions (Fig. 3.5a right).

We found that patterns of estimation biases and variability during numerosity estimation as a function of sensory evidence were in general more closely captured by the SEB relative to the TIM model (Fig. 3.5b top and middle panels). As predicted by our analytical analyses (Supplementary Note 3.4) the rate of increase in noise as a function of n is larger for the TIM model relative to the SEB model, with the empirical data more closely agreeing with the SEB model. Additionally, given that the TIM model generally requires larger values of lapse rates g to explain variance, for small n it predicts larger SDs relative to SEB and empirical data (with a similar pattern for the case of the CV, (Fig. 3.5a). A point where the TIM model appears to do a better job relative to the SEB model is for the absolute error estimations (Fig. 3.5b bottom). Subitizing is more pronounced for low numbers in general, and this reduces both biases and errors for $n < 5$. However, beyond the subitizing range and for levels of noise that challenge sensory perception, the SEB model does a better job at capturing all descriptive statistics. To visualize the nature of these differences, the posterior distribution of estimates for both models are shown in Fig. 3.5c for different numerosities and presentation times.

3.4. Discussion

We developed a model of efficient numerosity estimation based on Bayesian inference that endogenizes the environmental distribution and the sensory evidence (time or contrast) of the stimulus. Our theoretical and empirical tests provide clear evidence that a model of Bayesian decoding of noisy internal representations—which provides a normative explanation for the property of scalar variability and can be parsimoniously connected to a theory of limited informational capacity—provides a better account of numerosity estimation data in humans relative to the alternative TIM model considered here. We emphasize that both models: (i) are optimized for the same assumed objective (minimizing the MSE of the estimates), (ii) can be compared under the same assumption about the prior distribution, and (iii) have identical degrees of freedom. Thus, qualitative and quantitative differences between the two information-theoretical models cannot be explained by differences in model complexity, but instead reflect differences in the mechanistic assumptions of the numerosity processing operations. In particular, it is important to note that assumptions about potential encoding and decoding operations are explicitly stated in the Bayesian model. In contrast, these remain "hidden" in the alternative TIM model.

One of our main goals in the development of our modeling framework was to develop an encoding-decoding model incorporating various aspects of human cognition with many antecedents in the literature, which include Brownian motion during evidence processing over time [143] and logarithmic internal representation of numerical quantities [83]. While our proposed model accounts for key qualitative features of the human behavioral data with minimal degrees of freedom, we do not claim that the log-encoding model necessarily accounts for all aspects of numerosity estimation behavior. Indeed, the encoding and decoding strategies that humans and other animals use need not be the same in all contexts [156]. It is equally possible that numerosity processing mechanisms depend on the task at hand, and draw upon an ensemble of strategies that optimize performance under different situations [157, 158]. For instance, in future work, it will be interesting to investigate whether situations that involve explicit numerosity estimation vs discrimination rely on similar or distinct encoding strategies and inference processes.

We assumed that participants utilize a log-normal (or power-law) prior, however, it is important to note that the numerosities presented to the participants were drawn from a uniform distribution. We thus implicitly

assumed that participants did not rapidly adapt their encoding operations, which might be a reasonable assumption given that participants were not exposed to the new prior for an extended period of time. However, in one version of the model fits we allowed the parameters of the prior to be free parameters, resulting in non-uniform distributions. A natural consequence of our theory is that the SEB model parsimoniously endogenizes parameters of the prior distribution in its encoding operations. A testable prediction is that larger prior distribution ranges should lead to more noisy estimates and therefore poorer discriminability for a given capacity bound. This prediction is confirmed by a recent study where it is shown that human participants adapt their numerosity sensitivity for different numerosity ranges, with important implications for risk behaviour [159]. Thus two of the key predictions of our theory hold: for a fixed capacity bound sensory reliability should change as a function of (i) time exposure to the sensory stimulus as shown in this study, and (ii) the range of the prior distribution [159].

Additionally, our model predicts that the crossover point from overestimation to underestimation should change as a function of the numerosity range. In this work, we only present data with a fixed range of 1 to 15, thus we cannot test this prediction. However, a previous study using larger numerosity ranges (e.g., up to 30 or 100) found that the cross-over point is larger for wider numerosity ranges, and crucially, the degree of over- and under-estimation depended on the attentional resources dedicated to numerosity estimation [138]. This result is again in line with the predictions of our model.

Taken together, our findings suggest the fruitfulness of studying optimal models, which can serve as a departing point to understand the neuro-computational mechanisms underlying human behaviour without ignoring the fact that biological systems are limited in their capacity to process information [28, 71, 160, 161]. This highlights that understanding behavior in terms of its objectives while taking into account cognitive limitations, alongside encoding, decoding, and inference processes is likely to be essential to elucidate the mechanisms underlying human cognition.

3.5. Methods

Participants, data, and experiments

In this work we re-analyzed the data of experiments collected in previous work [50]. In brief, on each trial, between 1 and 15 dots were flashed, followed by a noise mask. The participants were then asked to type their guess of how

many dots were displayed. The participants were recruited and carried out the experiment online. There were three between-participant experiments (n=110 per experiment) that manipulated available stimulus information (variable exposure time: $t \in [40, 80, 160, 320, 640]$ ms) and different ways of controlling non-numerical properties of the stimuli: the average dot size (experiment 1), surface density (experiment 2) or surface area (experiment 3) of the dots.

We also studied a fourth experiment (n=110) in which time exposure t was fixed across trials, but instead display contrast of the dot arrays was varied from trial to trial (experiment 4). In this experiment, the colors of the dots varied between the background (grey) and pitch black, by Weber contrasts of 10%, 20%, 40%, 80% and 160%, at a constant presentation time of $t = 200$ ms.

Each participant was presented with each combination of numerosity and sensory evidence twice for a total of 150 trials per participant.

Models

Here we fit the two families of models described in the main text to the data of each participant: (i) We fit the SEB model assuming a log-normal prior with power parameter $\alpha = 2$. We fit a continuous version of the model based on the analytical solutions derived in Supplementary Notes 3.1-3.2, and a discrete version of this model based on numerical simulations. (ii) Following the procedures of previous work [50], we fit the TIM model assuming a power-law prior with power parameter $\alpha = 2$. For both families of models, we also fit a version were the parameters of the log-normal prior were allowed to be free parameters. We also note that analytical solutions in SEB were derived in the continuous space due to mathematical tractability (Supplementary Notes 3.1-3.3). Thus, in order to define the likelihood function of this model in the integer space, we normalized the log probability of estimators (Eq. 3.27) in the integer range $n \in [1, 2, 3, ..., 100]$. Note that both SEB and TIM have exactly the same degrees of freedom (R, B, and g), where g is a guessing rate based on the probability of randomly drawing a value from the default distribution.

Quantitative and qualitative analyses

Participants who completed less than 90% of the trials were excluded. Similar to previous work [50] we selected the 100 best participants for each experiment. In addition, trials in which the participant's response was 10 times higher

than the presented numerosity or the response time was superior to 10s were excluded. This additional data cleaning leads to the rejection of 142 trials out of 14,997 for experiment 1, 143 out of 14,993 for experiment 2, 172 out of 15,000 for experiment 3 and 187 out of 15,000 for experiment 4. Each model was fit individually to each participant using the DEoptim package [162] in the statistical language R [163] with a number of iterations set to 100. The limits for the parameter search space were set to (0.1,200) for R, (0.1, 20) for B and (0.0001, 0.5) for g. In the models where the prior was free, the search space of the prior parameters was (-50,50) for μ and (0.1,100) for σ. Model comparison was performed based on the Akaike information criterion (AIC). Using other model comparison metrics such as the Bayesian information criterion (BIC) does not change the conclusions of our work.

In Figs. 3.3- 3.4 and main text, we report the sum of the AIC difference relative to the best model across participants for each experiment, and report the 95% bootstrap confidence interval (95%-CI). We also computed two-sided paired t-tests based on the AICs obtained for each participant between the SEB and the TIM models. Likewise, we computed two-sided paired t-tests based on the guess rate parameter g obtained from each participant in the SEB model relative to the guess rates obtained in the TIM model. The qualitative predictions were computed based on the value with the highest density for each parameter at the population level. Each statistic was computed separately for each experiment and then averaged across experiments.

Details regarding the theoretical derivations of the SEB model and the analytical comparison between TIM and SEB models are given in detail in Supplementary Notes 3.1-3.4.

Acknowledgements

This work was supported by a European Research Council (ERC) starting grant (ENTRAINER) to R.P. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 758604), and support from the U.S. National Science Foundation, under grant SES-DRMS-1949418 to M.W.

Supplementary Tables

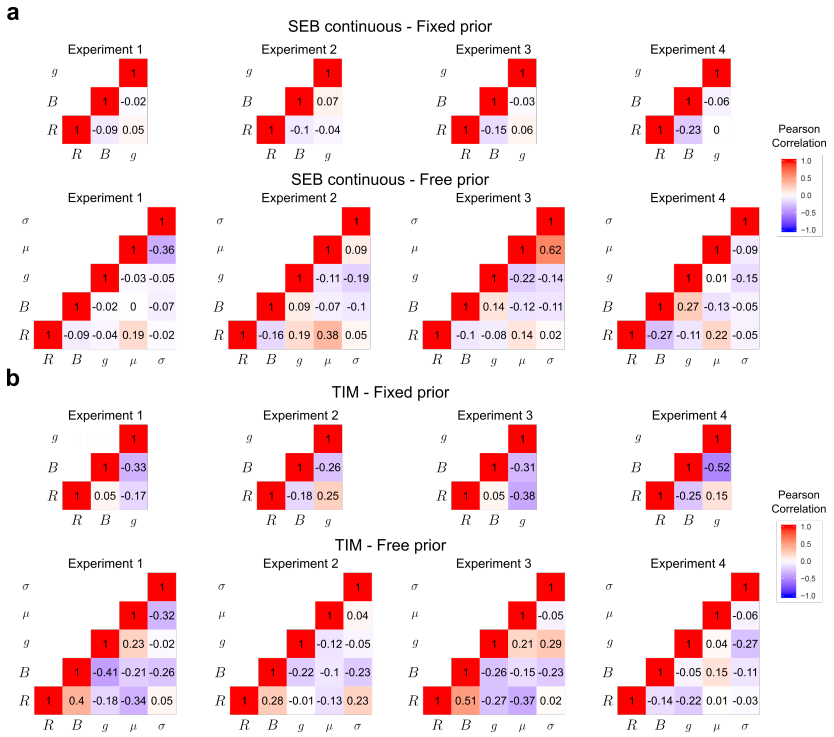
Prior parameters	Model	Experiment	R	B	g	μ	σ
Fixed	TIM	1	35.0	3.03	0.014		
		2	31.7	2.89	0.038		
		3	56.0	3.04	0.014		
		4	28.1	2.75	0.041		
	SEB discrete	1	13.4	7.09	0.010		
		2	10.6	8.68	0.012		
		3	13.7	7.62	0.014		
		4	8.5	6.71	0.029		
	SEB continuous	1	13.4	6.00	0.006		
		2	10.3	4.98	0.002		
		3	14.0	5.47	0.002		
		4	8.4	5.17	0.014		
Free	TIM	1	21.9	1.89	0.009	0.78	1.38
		2	20.1	1.73	0.010	1.35	1.74
		3	27.0	1.96	0.010	-1.17	2.51
		4	11.5	1.72	0.013	1.01	2.33
	SEB discrete	1	11.4	9.26	0.010	1.53	0.93
		2	10.7	8.04	0.009	1.43	0.99
		3	13.8	7.81	0.025	1.34	0.98
		4	7.7	9.24	0.013	1.00	1.19
	SEB continuous	1	13.9	6.18	0.010	1.29	0.64
		2	12.9	6.33	0.005	1.30	0.68
		3	14.7	5.52	0.004	1.42	0.77
		4	10.3	5.08	0.019	0.84	1.22

Supplementary Table 3.1. Parameter fits. Highest density of each parameter at the population level for each model and experiment.

Prior parameters	Model	Experiment 1	Experiment 2	Experiment 3	Experiment 4	All experiments
Fixed	TIM	55 136	63 016	56 126	68 453	242 731
	SEB discrete	54 660	59 113	55 318	64 796	233 887
	SEB continuous	**53 663**	**57 732**	**53 810**	**63 347**	**228 552**
Free	TIM	52 665	58 105	53 415	62 582	226 767
	SEB discrete	52 536	55 954	52 427	62 247	223 165
	SEB continuous	**50 978**	**54 420**	**50 486**	**60 444**	**216 328**

Supplementary Table 3.2. AIC of the model fit. The continuous version of SEB has the lowest AIC when the prior parameters are either fixed or free, which indicates a better fit to the behavioral data. In addition, the versions of the models with free prior parameters have lower AIC than the versions with fixed prior parameters.

Supplementary Figures



Supplementary Figure 3.1. Correlations of parameter fits. (a) Pearson correlation of the parameter fits of the SEB continuous for each experiment with fixed (top) and free (bottom) prior parameters. **(b)** Pearson correlation of the parameter fits of the TIM model for each experiment with fixed (top) and free (bottom) prior parameters. The parameters across participants are weekly correlated suggesting that they are identifiable as confirmed in parameter recovery exercises.

Supplementary Notes

Supplementary Note 3.1. General specification and derivation of the logarithmic noisy encoding and Bayesian decoding model

The goal of this supplementary note is to introduce the noisy log-encoding Bayesian model that was elaborated in more detail elsewhere [85]. Additionally, we formalize the connection between the log-normal and power-law priors. The definitions and derivations specified here will serve as a basis to specify the information theoretical part of the model developed in Supplementary Note 3.2.

In this model, we assume that a stimulus of numerosity n generates an internal representation that is drawn from a distribution

$$r \sim N(\log n, \nu^2), \tag{3.19}$$

where the noise parameter ν is independent of n. Here we assume that the prior distribution from which the numerosity value n is drawn is given by a log-normal distribution

$$\log n \sim N(\log \mu, \sigma^2). \tag{3.20}$$

As stated in the main text, this distribution is qualitatively similar to the power-law distribution and also has many occurrences and applications in the statistics of human behavior. It is also generally present in various biological phenomena such as measures of length, area and weight of living organisms, and also present in neurophysiological observations such as distribution of firing rates across populations of neurons and intrinsic gain and synaptic weight in neural systems [148].

Based on these two assumptions (Eqs. 3.19 and 3.20), it follows that the distribution of $\log n$ conditional on the value of r will be a Gaussian distribution

$$\log n \mid r \sim N\left(\mu_{\text{post}}\,(r), \sigma^2_{\text{post}}\right). \tag{3.21}$$

It follows that the conditional mean of $\log n$ is given by

$$\mu_{\text{post}}\,(r) = \mathrm{E}[\log n \mid r] = \mu + \beta \cdot (r - \mu), \tag{3.22}$$

with the slope of this linear projection given by

$$\beta = \frac{\sigma^2}{\sigma^2 + \nu^2}. \tag{3.23}$$

And the conditional variance is given by

$$\sigma^2_{\text{post}} = \frac{\sigma^2\nu^2}{\sigma^2+\nu^2}. \tag{3.24}$$

Here we consider the hypothesis that the participant's numerosity estimate minimizes the MSE. Thus, the rule that is optimal under this objective is given where the estimate $\hat{n}$ is defined by $\hat{n} = \mathrm{E}[n \mid r]$ for all r. It follows from the properties of the log-normal distribution that the posterior mean is given by

$$\mathrm{E}[n \mid r] = \exp\left(\mu_{\text{post}} + (1/2)\sigma^2_{\text{post}}\right). \tag{3.25}$$

In this case, the Bayesian model predicts

$$\begin{aligned}\log\hat{n}(r) &= \log \mathrm{E}[n \mid r] = \mu_{\text{post}}\,(r) + (1/2)\sigma^2_{\text{post}} \\ &= \mu + \beta\cdot(r-\mu) + (1/2)\sigma^2_{\text{post}}\end{aligned} \tag{3.26}$$

Given that r is a random variable, it follows that $\hat{n}(r)$ is also random variable. Thus, $\log\hat{n}$ is normally distributed conditional on n

$$\log\hat{n} \sim N\left(\hat{\mu}(n), \hat{\sigma}^2\right), \tag{3.27}$$

with the mean and variance of this conditional distribution given by

$$\begin{aligned}\hat{\mu}(n) &\equiv \mathrm{E}[\log\hat{n} \mid n] = \mu + \beta\cdot(\mathrm{E}[r \mid n] - \mu) + (1/2)\sigma^2_{\text{post}} \\ &= \mu + \beta\cdot(\log n - \mu) + (1/2)\sigma^2_{\text{post}} \\ \hat{\sigma}^2 &\equiv \mathrm{var}(\log\hat{n} \mid n) = \beta^2\,\mathrm{var}(r \mid n) \\ &= \beta^2\nu^2 = \frac{\sigma^4\nu^2}{\left(\sigma^2+\nu^2\right)^2}.\end{aligned} \tag{3.28}$$

It then follows from the properties of the log-normal distribution that the expected value and variance of the numerosity estimators are given by

$$\mathrm{E}[\hat{n} \mid n] = \exp\left(\hat{\mu}(n) + (1/2)\hat{\sigma}^2\right) \tag{3.29}$$

and

$$\mathrm{var}[\hat{n} \mid n] = \left[\exp\left(\hat{\sigma}^2\right) - 1\right]\cdot\exp\left(2\hat{\mu}(n) + \hat{\sigma}^2\right). \tag{3.30}$$

Finally, we can use the these equations to compute the ratio between the standard deviation and the expected value of the posterior estimators, i.e., the coefficient of variation (Eq. 3.12 in main text)

$$\frac{\mathrm{SD}[\hat{n} \mid n]}{\mathrm{E}[\hat{n} \mid n]} = \sqrt{e^{\hat{\sigma}^2} - 1} > 0.$$

This expression does not depend on n, and therefore the log-encoding Bayesian model delivers the property of *scalar variability* discussed in the main text.

Note that in these calculations, only the *normalized prior* $\tilde{p}(n) \equiv p(n)/p(1)$ matters, and in fact the Bayesian posteriors can be well-defined even in the case of an improper prior (for which $\tilde{p}(n)$ is well-defined, but there is no value for $p(1)$ such that the implied density function $p(n)$ will integrate to 1). All of the above calculations can be generalized to apply to any normalized prior of the form

$$\tilde{p}(n) \;=\; \exp\bigl(-\alpha(\log n) - \gamma(\log n)^2\bigr), \tag{3.31}$$

for some parameters α, γ with $\gamma \geq 0$. If $\gamma > 0$, this corresponds to a log-normal prior, with $\mu = (1-\alpha)/(2\gamma)$, $\sigma^2 = 1/(2\gamma)$. If instead $\gamma = 0$ but $\alpha > 0$, this corresponds to an improper power-law prior, $p(n) \sim n^{-\alpha}$.

In this latter case, the posterior implied by an internal representation r is again log-normal, as in equation (3.21), but now with parameters

$$\mu_{\text{post}}(r) \;=\; r \,+\, (1-\alpha)\nu^2, \qquad \sigma^2_{\text{post}} \;=\; \nu^2$$

as limiting cases of equations (3.22) and (3.24). It then follows that the Bayesian posterior mean estimate $\hat{n}(r)$ will be log-normally distributed conditional on the true value of n, as in equation (3.27), but with parameters

$$\hat{\mu}(n) \;=\; \log n \,+\, (\frac{3}{2} - \alpha)\nu^2, \qquad \hat{\sigma}^2 \;=\; \nu^2$$

as limiting cases of the formulas given above. Hence the mean estimate will be given by

$$\mathrm{E}[\hat{n}\,|n] \;=\; An, \qquad \text{where } A \equiv \exp\bigl((2-\alpha)\nu^2\bigr) \;>\; 0, \tag{3.32}$$

and the standard deviation of the estimates will again satisfy (3.12), with the value of $\hat{\sigma}^2$ given above.

Thus even in the case of an improper prior of this kind, the optimal Bayesian estimate $\hat{n}(r)$ is well-defined, and we can derive the predicted distribution of $\hat{n}$ conditional on n, as a function of the model parameters. All priors in the family (3.31) imply that the distribution of estimates should satisfy the property of scalar variability (3.12). In the case of a log-normal prior ($\gamma > 0$), equation (3.29) implies that $\mathrm{E}[\hat{n}\,|n]$ will be a strictly concave function of n, greater than n for all n below some critical value, and smaller than n for all

n above the critical value. In the limiting case of a power-law prior ($\gamma = 0$), instead, equation (3.32) implies that $\mathrm{E}[\hat{n}\,|n]$ should be proportional to n, with either overestimation for all n (if $\alpha < 2$) or underestimation for all n (if $\alpha > 2$). In the special case of a power law with $\alpha = 2$, the model implies that the optimal Bayesian estimate should be unbiased for all n.

Supplementary Note 3.2. Logarithmic noisy encoding and Bayesian decoding under limited informational capacity and temporal sensory exposure

In this supplementary note we show that it is possible to formulate an efficient coding model of numerosity estimation as developed in Supplementary Note 3.1, but in which encoding precision depends on stimulus viewing time t.

Instead of assuming, as in Supplementary Note 3.1, that a stimulus of numerosity n results in an internal representation r that is a single draw from a probability distribution that depends on n, we suppose now that the internal representation r instead consists of the sample path of a Brownian motion z_s over a time interval $0 \leq s \leq \tau$, starting from an initial value $z_0 = 0$. The drift m of the Brownian motion is assumed to depend on n, while its instantaneous variance ω^2 is independent of n; the length of time τ for which the Brownian motion evolves is also independent of n, but depends on the viewing time t. In assuming sensory evidence given by a Brownian motion with a drift that depends on the stimulus, we follow a long modeling tradition that includes the popular drift-diffusion model [143]. Models of this kind have been used since Taylor, Lindsey and Forbes [151] to account quantitatively for the way in which the accuracy of perceptual judgments is affected by manipulations of viewing time.

More specifically, we assume that m is an affine transformation of the logarithm of n,

$$m \;=\; \xi + \psi \log n, \tag{3.33}$$

where the parameters ξ and ψ may depend of the statistics of a particular environment. We suppose that the choice of these coefficients is subject to a "power constraint" which requires the average value of m^2 to be within some finite bound

$$\mathrm{E}[m^2] \;\leq\; \Omega^2 \;<\; \infty. \tag{3.34}$$

This bound on the amount of variation in the drift limits the precision with which different stimuli can be discriminated, for any given τ. The value of τ is assumed to grow linearly with the viewing time, up to some finite bound B,

$$\tau \;=\; \min(t,\, B), \tag{3.35}$$

representing a constraint on the amount of evidence that can be maintained in working memory. The latter bound constrains the degree to which precision can be increased by further increases in viewing time, just as in the TIM model.

For any fixed value of τ, the final position z_τ of the Brownian motion at time τ is a sufficient statistic for the information contained in the sample path about the value of n. Hence Bayesian decoding of the information contained in the sample path will yield the same result as if the internal representation is assumed simply to be the scalar random variable z_τ, with distribution

$$z_\tau \sim N(m(n)\tau, \omega^2\tau). \tag{3.36}$$

Alternatively, we may suppose that the internal representation of n is given by the scalar random variable $r \equiv z_\tau/\tau$, which contains the same information as the variable z_τ. Under this representation of the sensory evidence, r is a draw from a distribution

$$r \sim N(m(n), \omega^2/\tau). \tag{3.37}$$

Equation (3.37) effectively states that r is the output of a *Gaussian channel* with input m [152]; hence the problem of optimally choosing the function $m(n)$ is equivalent to an *optimal encoding* problem for a Gaussian channel. The capacity C of such a channel is a quantitative upper bound on the amount of information that can be transmitted regardless of the encoding rule, which is equal to

$$C = \frac{1}{2}\log\left(1 + \frac{\Omega^2 t}{\omega^2}\right), \tag{3.38}$$

an increasing function of Ω/ω as well as of t. Here we suppose that the goal is to design a system that minimizes the mean squared error of the estimate $\hat{n}$ when n is drawn from a log-normal prior distribution (Eq. 3.20).

$$\log n \sim N(\log \mu, \sigma^2).$$

Note that the estimate $\hat{n}$ depends on r_n. We re-express r_n as a function of the transformed variable $\tilde{r}_n \equiv (r_n - \xi)/\psi$, thus we can equivalently treat $\tilde{r}_n$ as the internal representation, and it follows that $\tilde{r}_n \sim N(\log n, \omega^2/(t\psi^2))$. Following the definitions provided in Supplementary Note 3.1, it follows that we have a noisy log-encoding with variance $\nu^2 = \omega^2/(t\psi^2)$. It then follows

that the MSE in the case of any encoding rule is given by Eq. 3.26, and the MSE associated with this rule will be given by

$$\mathrm{MSE} = \exp\left(2\mu + 2\sigma^2\right) \cdot \left[1 - \exp\left(-(1-\beta)\sigma^2\right)\right]. \tag{3.39}$$

Recall that β is a decreasing function of ν (Eq. 3.23), and therefore in order to make the MSE as small as possible it is desirible to make ν as small as possible. Given that $\nu^2 = \omega^2/(t\psi^2)$ is follows that we would like to make ψ as large as possible, consistent with the power constraint in Eq. 3.34. Thus, for the case of the log-normal prior the power constraint becomes

$$(\xi + \psi\mu)^2 + \psi^2\sigma^2 \leq \Omega^2. \tag{3.40}$$

The maximum value of ψ consistent with this constraint is achieved when

$$\xi = -\psi\mu, \quad \psi = \frac{\Omega}{\sigma}. \tag{3.41}$$

In this case the encoding noise is given by

$$\nu = \frac{\omega}{\Omega\sqrt{t}}\sigma. \tag{3.42}$$

Defining $R \equiv \Omega/\omega$, we can define the encoding noise of numerosity estimation

$$\nu(t) = 1/G, \tag{3.43}$$

where

$$G = \min(R\sqrt{t}/\sigma, B), \tag{3.44}$$

with B a maximum biologically allowed bound on sensory precision (similar to the assumption of the TIM model).

The precision of numerosity encoding is given by $\nu(t) = 1/G$, where $G = \min(R\sqrt{t}/\sigma, B)$ and B a maximum biologically allowed bound on sensory precision.

Supplementary Note 3.3. Recapitulation of the Gaussian channel capacity derivation

For convenience to the reader, the goal of this supplementary note is to recapitulate the derivation of the Gaussian channel capacity presented in Cover and Thomas [152] based on the notation used in our work, thus clarifying the connection to the solution of our SEB model under capacity constraints derived in Supplementary Note 3.2.

Suppose that Y is the output of a channel with input $X + Z$, where X is the signal and Z the noise. We assume the noise is drawn from a Gaussian distribution with variance ω^2/t and mean 0.

The goal is to find the maximum achievable channel capacity C by maximizing the mutual information $I(X;Y)$ for a given power constraint Ω^2

$$C = \max_{f(x):EX^2 \leq \Omega^2} I(X;Y) \tag{3.45}$$

It can be shown that

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y \mid X) \\ &= h(Y) - h(X + Z \mid X) \\ &= h(Y) - h(Z \mid X) \\ &= h(Y) - h(Z), \end{aligned} \tag{3.46}$$

where in general $h(X)$ is defined as the entropy of the channel X. Here, we will use two results. First, the entropy of a Gaussian channel, say channel for Z with given variance ω^2, is given by

$$h(Z) = \frac{1}{2} \log 2\pi e \omega^2/t. \tag{3.47}$$

Second,

$$EY^2 = E(X+Z)^2 = EX^2 + 2EXEZ + EZ^2 = \Omega^2 + \omega^2/t, \tag{3.48}$$

given that X and Z are independent and $EZ = 0$. This means that the entropy of Y is bounded by

$$h(Y) \leq \frac{1}{2} \log 2\pi e(\Omega^2 + \omega^2/t). \tag{3.49}$$

Thus, replacing Eqs. 3.47 and 3.49 in Eq. 3.46 gives

$$\begin{aligned} I(X;Y) &= h(Y) - h(Z) \\ &\leq \frac{1}{2}\log 2\pi e(\Omega^2 + \omega^2/t) - \frac{1}{2}\log 2\pi e\omega^2/t \\ &= \frac{1}{2}\log\left(1 + \frac{\Omega^2 t}{\omega^2}\right). \end{aligned} \tag{3.50}$$

Supplementary Note 3.4. Comparison between the family TIM and SEB models

In this supplementary note, we compare the kind of information theoretical model that does not incorporate Bayesian inference (the TIM model defined in the main text), and the general family of Bayesian encoding-decoding models. This illustrative comparison is not directly applicable to our numerosity estimations, but can be solved analytically and is instructive. Nevertheless, we provide numerical simulations that are applicable to numerosity estimation in the main text which confirm the main predictions presented in this note (see Fig. 3.2 in main text).

Illustrative comparison between the family of TIM and SEB models

In this illustrative example, the two models can be solved analytically, thus allowing to highlight the commonalities and differences between both models in an intuitive manner.

The TIM model proposes a method to infer how the distribution of estimates $\hat{n}$ should vary depending on the true stimulus magnitude n. The goal is to find the response distribution $p(\hat{n}|n)$ that minimizes the mean squared error (MSE)

$$\text{MSE} \equiv \int_N p(\hat{n} \mid n)(\hat{n} - n)^2 d\hat{n} \tag{3.51}$$

subject to the constraint that (Eq. 3.16 in main text)

$$D_{KL}\left(p_n \| q\right) \leq C(t)$$

where p_n is the distribution of possible responses conditional on n, q is the "prior" distribution, $D_{KL}\left(p\|q\right)$ is the Kullback-Leibler divergence, and $C(t)$ is a positive bound that depends on the amount of time t for which the stimulus is presented.

The TIM model can be developed further by specifying that q is given by the prior distribution from which n is expected to be drawn, and that $C(t)$ increases linearly with time, i.e., that $C(t) = c \cdot t$ for some $c > 0$, up to some finite bound B. Assuming that the prior is known, the model thus has only a single free parameter (the value of c) to predict the distribution of responses, as a function of both n and t, for all values of t below some upper bound.

We compare the predictions of this kind of model to the alternative Bayesian model, according to which (i) estimates are based on a noisy internal representation r of the stimulus magnitude n, which consists of a sequence of independent draws of a signal, the distribution of which depends on n, and with the number of draws in the sequence growing with t; and (ii) given the noisy internal representation, the participant's estimate is given by $\hat{n}(r) = \mathrm{E}[n \mid r]$: Note that the computation of this last conditional expectation must be relative to a particular prior distribution from which n is expected to be drawn. Given the distribution of possible samples r for each n and t, we can use the assumed response rule to derive a predicted distribution of responses $\hat{n}$ for any specification of (n, t).

We make the Bayesian model example more specific by assuming that r is the cumulative value at time t of a Brownian motion that starts from the initial value $r = 0$, with a drift $m(r)$ that depends on the stimulus and an instantaneous variance ω^2 that is independent of the stimulus. If we further assume that $m(n) = \mu \cdot n$ for some $\mu > 0$; then the model's predictions depend only on a single parameter, the value of $\gamma \equiv \mu/\omega$; again assuming that the prior is known. We thus have two one-parameter models, each of which makes precise predictions for the distribution $p_t(\hat{n} \mid n)$ for any n and t. Thus, in each model, the single free parameter determines how rapidly the precision of estimates should improve with increasing viewing time.

In this example, we suppose that the prior distribution for n is Gaussian, and let it be given by $N(0, \sigma^2)$. Here, we economize in notation by assuming that the prior mean is zero; the formulas that follow hold regardless of this, but n should be understood as the stimulus magnitude relative to the prior mean, and $\hat{n}$ as the response relative to the prior mean.

In the case of the Bayesian model, the information contained in the noisy internal representation is equivalent to that for a model in which the available information is a noisy measurement, $r \sim N(n, (\gamma^2 t)^{-1})$, the precision of which grows linearly with t. The optimal Bayesian estimate is then $\hat{n}(r) = \phi_t \cdot r$, where

$$\phi_t \equiv \frac{\sigma^2}{\sigma^2 + (\gamma^2 t)^{-1}}. \tag{3.52}$$

From this, it follows that the conditional distribution of responses for any time t will be given by

$$\hat{n} \mid n, t \sim N\left(\phi_t n, \phi_t \left(1 - \phi_t\right) \sigma^2\right). \tag{3.53}$$

For the TIM model instead: the first order conditions for minimization of Eq. 3.17 subject to Eq. 3.16 require that

$$(\hat{n} - n)^2 + \theta_{n,t} \ln \frac{p(\hat{n} \mid n)}{q(\hat{n})} = k_{n,t}. \tag{3.54}$$

for all $\hat{n}$, where $\theta_{n,t}$ is the Lagrange multiplier associated with the capacity constraint given in Eq. 3.16 for given choices of n and t, and $k_{n,t}$ is a constant of integration. This equation can be solved for $p(\hat{n}|n)$ for each $\hat{n}$, given values for $k_{n,t}$ and $\theta_{n,t}$. We choose $k_{n,t}$ so as to imply a PDF $p(\hat{n}|n)$ that integrates to 1. We see that the resulting distribution for $\hat{n}$ is Gaussian

$$\hat{n} \mid n, t \sim N\left(\phi_{n,t}, (1 - \phi_{n,t})\, \sigma^2\right), \tag{3.55}$$

where the bias coefficient $\phi_{n,t}$ corresponds to

$$\phi_{n,t} = \frac{2\sigma^2}{\theta_{n,t} + 2\sigma^2}. \tag{3.56}$$

The value of $\theta_{n,t}$ is chosen so as to imply that the constraint given in Eq. 3.16 holds with equality. Computing the KL divergence, we see that this holds if and only if

$$\Gamma\left(\phi_{n,t}\right) + \phi_{n,t}^2 \frac{n^2}{\sigma^2} = 2C(t), \tag{3.57}$$

where

$$\Gamma(\phi) \equiv -\ln(1 - \phi) - \phi \tag{3.58}$$

for any $0 < \phi < 1$.

For any n, we note that $\Gamma(\phi)$ is a continuous, monotonically increasing function of ϕ, approaching 0 as $\phi \to 0$ and becoming unboundedly large as $\phi \to 1$. Hence, for any n and any $C(t) > 0$, equation 3.57 has a unique solution satisfying $0 < \phi_{n,t} < 1$. We further observe that for a fixed value of n, increasing $C(t)$ increases the value of $\phi_{n,t}$; and for a fixed $C(t)$, increasing the value of $|n|$ increases the value of $\phi_{n,t}$.

Based on these analytical solutions, these results reveal some important similarities between the predictions of the TIM and the Bayesian models: for large enough t and allowing $C(t)$ to grow as function of t, then both models imply

1. $\mathrm{E}[\hat{n} \mid n] \to n$, and
2. $\mathrm{var}[\hat{n} \mid n] \to 0$.

Nonetheless, there are also several notable differences in the predictions of the two models. Here we provide a detailed explanation of these differences which were already mentioned in the main text:

1. *The quantitative dependence of estimation bias on viewing time.* While both models predict that $\phi_{n,t}$ should increase from 0 (for $t = 0$) to 1 (as $t \to \infty$), they do not imply the same rate of increase in $\phi_{n,t}$ as t increases. The Bayesian model implies that

$$\frac{\phi_t}{1-\phi_t} = \sigma^2\gamma^2 t, \tag{3.59}$$

for any n. Hence for small t, $\phi_t \sim t$, while for large t, $(1-\phi)^{-1} \sim t$. Instead, if in the TIM model we assume that $C(t) = c \cdot t$ for all t, then for any $n \neq 0$, one can show that the solution to Eq. 3.57 satisfies $\phi_{n,t} \sim t^{1/2}$ for small t, while $(1-\phi)^{-1} \sim e^{2ct}$ for large t. Thus regardless of the parameters γ and c for the two models, we see that the TIM model implies faster growth of $\phi_{n,t}$ as t increases, both for sufficiently small values of t and for sufficiently large values of t.

2. *The relationship between estimation bias and the variability of estimates.* Again fixing some single value of $n \neq 0$, and considering the implied distribution of estimates for different viewing times, we see that the two models do not imply that $\mathrm{var}[\hat{n}|n,t]$ co-varies with the bias in the same way. The TIM model implies that the variance falls monotonically with increases in $\phi_{n,t}$ (and hence that the variance falls monotonically with time, for any n). The Bayesian model instead implies that increases in ϕ_t first increase variance (while ϕ remains below 1/2), and then reduce variance again (once $\phi_t > 1/2$). The difference in predictions is especially stark in the case of small viewing time. As $t \to 0$, the Bayesian model implies that the variance should fall to zero (estimates are based on the expected value of the prior), while the TIM model implies that this is the case in which estimates should be most variable (estimates are simply samples from the prior distribution, regardless of the value of n).

There are other differences between the two models that we do not highlight here as they are not strictly relevant to the discussion of this article.

Taken together, we developed an example in which analytical analyses allowed us to examine commonalities and differences between the two models. While the exact predictions of these differences do not hold for the specific application of the numerosity estimation models developed for the TIM model and

the noisy log-encoding Bayesian model (see Supplementary Notes 3.1 and 3.2), these general differences make the two numerosity models identifiable, and thus generate different qualitative predictions. In particular, the two differences highlighted above cause the TIM model not to provide a general account of the scalar variability principle. That is, the ratio between variability and expected value estimations grows more rapidly in the TIM model relative to the log-encoding Bayesian model.

4

CAUSAL PHASE-DEPENDENT CONTROL OF NON-SPATIAL ATTENTION IN HUMAN PREFRONTAL CORTEX

J. Brus*, J. A. Heng*, V. Beliaeva, F. Gonzalez, A. M. Cassará, E. Neufeld, M. Grueschow, R. Polanía. Causal phase-dependent control of non-spatial attention in human prefrontal cortex. In review.

Contributions

Conceptualization, Experimental software design, EEG data analysis, Writing

4.1. Abstract

Non-spatial attention is a fundamental cognitive mechanism allowing organisms to orient the focus of conscious awareness toward sensory information that is relevant to a behavioral goal while shifting it away from irrelevant stimuli. It has been suggested that attention is regulated by the ongoing phase of slow excitability fluctuations of neural activity in the prefrontal cortex, a hypothesis that has been challenged with no consensus. Here, we developed a behavioral and non-invasive stimulation paradigm aiming at modulating slow excitability fluctuations of the inferior frontal junction, and show that non-spatial attention can selectively be modulated as a function of the ongoing phase of exogenously modulated excitability states of this brain structure. These results demonstrate that non-spatial attention relies on ongoing prefrontal excitability states, which are likely regulated by slow oscillatory dynamics, that orchestrate goal-oriented behavior.

* These authors contributed equally to this work

4.2. Introduction

Is that a predator behind the bush? Is it moving to the left or to the right? When part of these spatially overlapping sensory features are more relevant to guide behavior than others, activity in sensory areas representing properties of the attended features are enhanced [164]. This cognitive process is known as non-spatial attention, allowing organisms to orient the focus of conscious awareness toward sensory information that is relevant to a behavioral goal while shifting it away from irrelevant stimuli. Non-spatial attention is commonly subdivided into feature-based attention, focusing on one single feature (such as a direction of motion, color or orientation) and object-based attention in which a participant has to attend a combination of features (such as an object or a scenery). There is consensus that this process is not an intrinsic property of sensory areas but relies on long-range functional interactions with prefrontal structures. While a large body of work implicates the inferior frontal junction (IFJ) as a key source of control signals for both forms of top-down non-spatial attention [51, 165, 166], the causal mechanisms of top-down regulation of attentional control remain unclear.

Based on behavioral observations that attentional performance fluctuates over time, rhythmic control has been proposed as a candidate mechanism of attentional regulation [167–169]. Supporting this notion, a study showed that temporal dynamics of attentional behavior closely resemble the spectral features of ongoing oscillatory brain activity in prefrontal structures [170]. Therefore, it was hypothesized that relatively slow and periodic neuronal excitability fluctuations might shape attention and overt behavior. However, the conclusions from many of these studies have been called into question, by suggesting that previously reported rhythmic variations of attentional behavior might be artifacts of the analysis approaches [171]. Moreover, whether ongoing excitability states within prefrontal structures are causally involved in regulating non-spatial attention, remains unknown.

Here, we attempt to reconcile some of these concerns using a behavioral paradigm coupled with a non-invasive brain stimulation protocol aiming at modulating, with high temporal precision, excitability fluctuations in the IFJ during non-spatial attention in the intact human brain. Here, it is important to emphasize that in our work we do not study the role of endogenous oscillatory fluctuations, but instead study the causal involvement of ongoing excitability states likely driven by slow rhythmic fluctuations (which in our case are exogenously controlled) on top-down attention. It is important to highlight that the causal involvement of prefrontal structures

during non-spatial attention has been demonstrated in previous landmark studies using transcranial magnetic stimulation (TMS) [165]. However, TMS induces only transient disruptions of neural functioning leaving the role of top-down control through slow fluctuations of the excitability state in prefrontal structures unresolved.

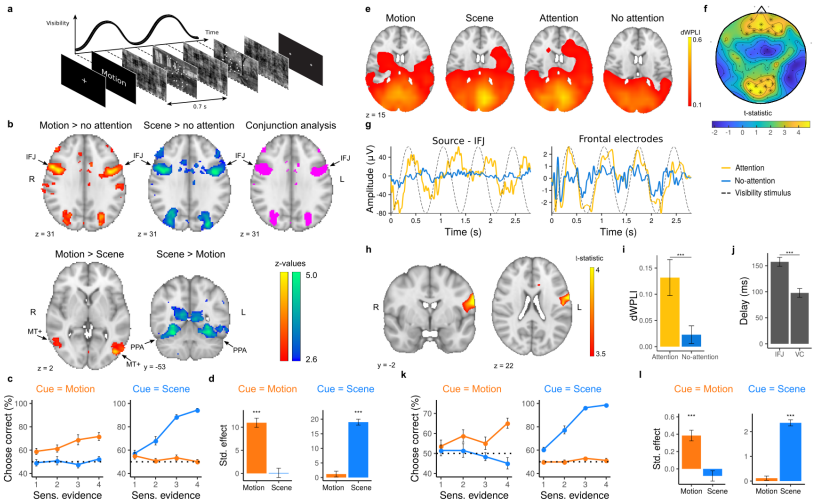


Figure 4.1. fMRI and EEG paradigm, Experiments 1 and 2. a) Example display of one trial. After the attentional cue, a sequence of four to seven compound stimuli is presented following a sinusoidal rhythm through time at 1.43 Hz. Participants respond with a button press, only taking the last motion/scene stimulus into account. If motion was cued participants press left for leftward motion and right for rightward motion, if scenes were cued they press left for indoor and right for outdoor scenes. **b)** fMRI results, Experiment 1. Attention to motion and scenes vs no-attention show that the inferior frontal junction (IFJ) activates bilaterally. A contrast of attention to motion vs scenes shows that the area associated with motion perception, the middle temporal complex (MT+), activates. The inverse contrast shows that the area sensitive to scene recognition, the parahippocampal place area (PPA), activates. Images were thresholded at $Z > 2.6$ and whole-brain cluster corrected, $P < 0.05$. **c)** Behavioral results, Experiment 1. Participants use the motion evidence when cued to pay attention to motion (orange) and the scene evidence when cued for scene (blue) and crucially ignore the irrelevant sensory feature. Error bars denote ± SEM.

Figure 4.1:
(continued) **d)** Standardized coefficients of a multifactor logistic regression of task performance as a function of evidence levels show that the participants are significantly influenced by the cued evidence (cue = motion $\beta_{\mathrm{RFX}} = 11.0$, $P_{\mathrm{MCMC}} < 0.001$, cue = scene $\beta_{\mathrm{RFX}} = 19.0$, $P_{\mathrm{MCMC}} < 0.001$) and not distracted by the irrelevant sensory feature. The standardized effect represents the expected value of the corresponding posterior beta estimate divided by its standard deviation. **e)** EEG results, Experiment 2. The dWPLI between the EEG data during the four first periods of the visual stimulus and the 1.43 Hz visual signal is computed. The dWPLI values show that a wide area of the visual cortex gets tagged to the frequency of the visual stimulation. **f)** The statistical difference in dWPLI between the attention and the no-attention task at the sensor level. Starred electrodes represent significant electrodes (cluster corrected at $P < 0.01$). **g)** Event-related potential for the first four periods of the visual stimulus of an example participant at the source (left) levels and sensor (right level). The source level signal corresponds to an IFJ voxel and the sensor level signal is shown for the frontal cluster left of electrodes with a higher debiased weighted phase lag index (dWPLI) during the attention vs no-attention task (see panel f). **h)** The dWPLI between the beamformed signals of each voxel and the visual stimulus was computed for attention vs no-attention. Maps show the statistical difference between the two attention conditions revealing the left IFJ to be tagged to the degree of stimulus visibility during attention trials (whole-brain cluster corrected at $P < 0.01$). **i)** dWPLI values of the IFJ cluster in panel g. **j)** The activation delay after visual stimulus presentation was larger for prefrontal vs visual cortex (VC) $T(18) = 5.06$, $P < 0.001$. Error bars denote ± SEM. **k)** and **j)** Task performance and standardized coefficients of a multifactor logistic regression of task performance in Experiment 2 replicate the effects observed in Experiment 1 (see panels **l** and **d**).

4.3. Results

Spatial and dynamic characterization for the modulation of non-spatial attention

We designed a behavioral paradigm with the primary goal of inducing a tagged oscillation in the IFJ during non-spatial attention, which would allow us to implement a closed-loop-like simulation protocol to modulate ongoing IFJ excitability states. Participants viewed two spatially overlapping sensory stimuli: (i) a cloud of dots from which a proportion was moving coherently to the left or right side of the screen and (ii) images of indoor or outdoor scenes. A series of stimuli went in and out of "phase coherence" in a sinusoidal manner (at 1.43 Hz) so that they were modulated in visibility over time while assuring that changes in luminance and spectral power remained constant (Figure 1a, Methods). In each trial, participants were cued to attend one of the two sensory features. At the end of each stimuli stream, participants were asked to indicate whether the last observed cloud of dots was mainly moving to the left or right (motion cue), or whether the last observed scene was indoor or outdoor (scene cue). The level of sensory evidence in the

last stimulus was randomly chosen from one of four predefined levels, thus, allowing us to modulate task difficulty trial-by-trial, where the smaller the sensory evidence the more difficult the trial (Methods). First, we made use of both fMRI (Experiment 1) and high-density EEG (Experiment 2) to investigate and validate both the spatial and dynamic involvement of the IFJ in our non-spatial attention task. Crucially, we implemented a control "no-attention" task that contained identical visual input as the non-spatial attention task, but where the stream of fluctuating sensory information was behaviorally irrelevant (Supplementary Figure 4.1, Methods).

In Experiment 1, we found that the bilateral IFJ was the most active brain area in the attention task compared to the no-attention task for each sensory modality (peak $Z_{\text{motion}} = 5.9$, $Z_{\text{scene}} = 6.1$, $P < 0.001$, $P < 0.05$ cluster corrected, Figure 1b), with a high degree of overlap across the two sensory modalities (conjunction analysis $Z > 2.6$, $P < 0.05$ cluster corrected, Figure 1b). The contrast of attention to motion vs scene showed that the bilateral middle temporal complex (MT+) was selectively active during motion-cued trials (peak $Z = 5.7$, $P < 0.001$, $P < 0.05$ cluster corrected, Figure 1b), and this result was accompanied by significant psychometric performance for motion evidence ($\beta_{\text{RFX}} = 11.0$, $P_{\text{MCMC}} < 0.001$, Figure 1c, d), but not for scene evidence ($\beta_{\text{RFX}} = 0.1$, $P_{\text{MCMC}} = 0.11$). On the other hand, the parahippocampal place area (PPA) was more active during scene cue trials (peak $Z = 4.6$, $P < 0.05$ cluster corrected, Figure 1b), and this result was accompanied by significant psychometric performance for scene evidence ($\beta_{\text{RFX}} = 19.0$, $P_{\text{MCMC}} < 0.001$, Figure 1c, d), but not for motion evidence ($\beta_{\text{RFX}} = 1.2$, $P_{\text{MCMC}} = 0.45$). As a sanity check, we show the main effects of the task (i.e., without contrasting attentive vs non-attentive states) and find that most of the visual cortex is active both when paying attention to motion and scenes (Supplementary Figure 4.2a), thus suggesting the specificity of top-down control involving the fonto-parietal network which prominently engages the IFJ (Figure 1b).

Next, we investigated whether the IFJ was indeed rhythmically tagged to the stimulus visibility, and if so, would this be more prominent during attention vs no-attention. First, we show raw dWPLI values, without contrasting attentive vs non-attentive states, and find, as expected, that most of the visual cortex gets entrained to the frequency of the visual input, Figure 2e. Scalp EEG analyses revealed clusters where the phase consistency between EEG signals and the tagged signal was higher in the attention vs no-attention condition ($T_{\text{max}} = 4.81$, $P_{\text{MCMC}} < 0.001$, $P < 0.01$ whole-brain cluster corrected, Figure 1f). Performing the same analyses at the source level, we found that

the most prominent significant cluster was located at the level of the left IFJ ($T_{\max} = 5.51$, $P_{\mathrm{MCMC}} < 0.001$, $P < 0.01$ cluster corrected, Figure 1h, i). To estimate the latency of sensory responses in the IFJ during the attention task, we extracted the relative phase between the frequency-tagged response and the stimulus on the screen. The average phase lag of the IFJ was 150 ms which was shifted by about 50 ms in comparison to early sensory areas ($T(18) = 5.06$, $P < 0.001$, Figure 1j), likely related to synaptic delays between areas, and roughly following previous reports [51]. At the behavioral level, these results were accompanied by a significant impact of motion-evidence on performance when motion was cued ($\beta_{\mathrm{RFX}} = 0.38$, $P_{\mathrm{MCMC}} < 0.001$, Figure 1k, l), but not for scene evidence ($\beta_{\mathrm{RFX}} = -0.08$, $P_{\mathrm{MCMC}} = 0.08$). Conversely, when scene was cued, psychometric performance was significant for scene evidence ($\beta_{\mathrm{RFX}} = 2.36$, $P_{\mathrm{MCMC}} < 0.001$, Figure 1k, l), but not for motion evidence ($\beta_{\mathrm{RFX}} = 0.12$, $P_{\mathrm{MCMC}} = 0.09$). Taken together, our set of neuroimaging results confirm that the IFJ gets tagged to the rhythmic stimulus presentation in our non-spatial attention task.

Additionally, we investigated whether some of the above-mentioned differences in top-down attentional control by the IFJ could be related to stronger oculomotor engagement in our task. Analyses of eye tracking data show that there is no significant difference between the number of microsaccades in either the motion, scene, or no-attention condition, therefore differences in eye movements can not explain the differences in brain activity (Supplementary Figure 4.2b). These results confirm the involvement of the IFJ during non-spatial attention in our task and the selectivity of sensory areas for each relevant feature.

Exogenous slow fluctuations modulation of IFJ top-down control

Having established rhythmic IFJ engagement during non-spatial attention in our task, the fundamental question we ask now is whether the slow fluctuations of the excitability state exogenously induced in the IFJ are causally related to top-down control. A key feature of our behavioral paradigm is that it allows us to predict latencies at which neural excitability for sensory processing is high. We hypothesized that boosting periods of predicted high excitability states in the IFJ would promote perceptual discriminability performance for the cued sensory feature. Conversely, downregulating periods of predicted high excitability states would hinder behavioral performance (Figure 2b). To test this hypothesis, we employed transcranial alternating current stimulation (tACS), a technique that has the potential to establish a causal link between oscillatory patterns—modulated or induced [56, 172–176]—at the targeted

brain structure and the resulting behavior. We applied tACS targeting the IFJ bilaterally using a ring electrode configuration to increase the focality of the induced electric fields (Figure 2a, Supplementary Figure 4.4, Methods). We applied 5% EMLA cream under the stimulation electrodes, allowing us to reduce somatosensory effects, increase stimulation intensities (up to 4 mA peak-to-peak, Methods), and thereby increase the chances of oscillatory neuromodulation. tACS was applied at the same sensory tagging frequency (1.43 Hz), but crucially, the presentation of sensory stimuli was precisely synchronized to the tACS waveform in one of two ways in each trial: first, the peak of anodal stimulation of the center electrode (defined as the peak of the waveform) coincides with periods of high sensory excitability ("in-phase" condition, while considering the delays estimated in the EEG experiment, Methods), which we expect to result in attentional improvements because anodal stimulation is thought to increase the excitability states of the targeted cortical structure [177]. Second, the peak of cathodal stimulation of the center electrode (which we define as the trough of the waveform) coincides with periods of high sensory excitability ("out-of-phase" condition), which should result in attentional hindering by reducing the cortical excitability states of the IFJ [177] (Figure 2b, Supplementary Figure 4.3).

In one of two lab visits, participants received "in-phase" tACS for one of the two sensory cues (attend to motion or scene, Experiment 3a) and received "out-of-phase" for the other sensory cue. The stimulation conditions were switched for each sensory cue in the second lab visit (Experiment 3b, Methods). We first investigated whether, during the "stimulation on" trials, in-phase stimulation improved behavioral performance relative to out-of-phase stimulation. In line with our hypothesis, we found that, across different sessions, in-phase stimulation improved sensory discrimination performance when motion was cued (interaction sensory evidence*stimulation-condition $\beta_{\mathrm{RFX}} = 2.8$, $P_{\mathrm{MCMC}} = 0.004$, Figure 2e), however, we did not find a significant effect when scenes were cued ($\beta_{\mathrm{RFX}} = 1.3$, $P_{\mathrm{MCMC}} = 0.086$). Post hoc analyses revealed that discrimination performance improved in the hypothesized direction for motion discrimination at the highest levels of difficulty (paired samples Wilcoxon test $P = 0.0014$ and $P = 0.013$ for levels 1 and 2, respectively, Figure 2d) and for scene discrimination at the highest level of evidence ($P = 0.018$). We employed the same multi-factor regression to investigate whether stimulation exerted influences on the distractor (non-cued) sensory feature. We found no effect of stimulation in either task ($P_{\mathrm{MCMC}} > 0.16$ in both tasks, Figure 2e). This indicates that modulations

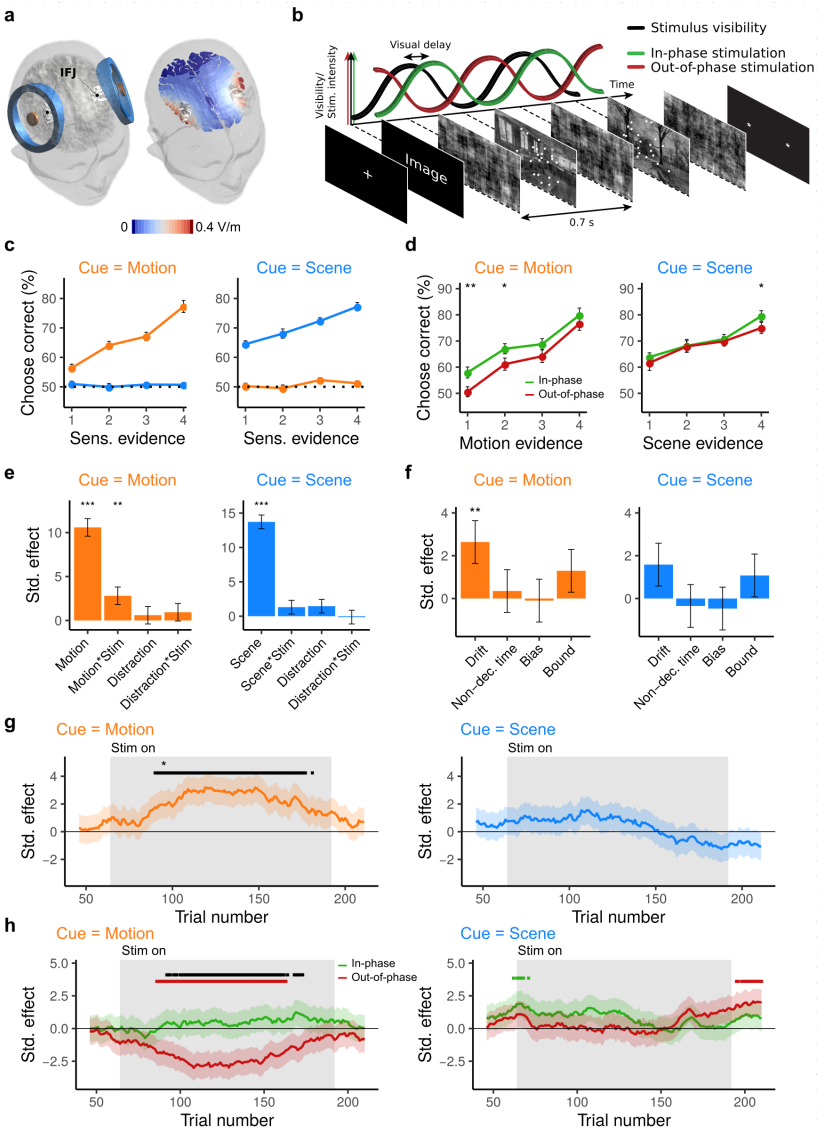


Figure 4.2. Temporal alignment of tACS over the IFJ modulates sensory perception, Experiment 3. a) Two concentric electrode pairs are placed over the left and right IFJ reaching relatively focused peak electric fields ~ 0.5 V/m (see Supplementary Figure 4.4). **b)** The tACS current follows a sinusoidal function applied either "in-phase" relative to the visual tagging response, or "out-of-phase" with a phase lag of 180° relative to the visual tagging response. **c)** The percentage of correct trials at different difficulty levels shows that participants use the cued sensory evidence and ignore the irrelevant ones.

Figure 4.2:
(continued) **d)** Participants perform better in the "in-phase" tACS condition compared to "out-of-phase" when they are cued to pay attention to motion mostly at the hardest difficulty levels (paired samples Wilcoxon test $P = 0.0014$ and $P = 0.013$ for levels 1 and 2, respectively) and for scene discrimination at the highest level of evidence ($P = 0.018$). **e)** A linear mixed-effects model reveals that in the motion trials (besides the main effect of motion evidence, $\beta_{\mathrm{RFX}} = 10.6$, $P_{\mathrm{MCMC}} < 0.001$) there is a significant interaction effect between motion evidence and stimulation condition ($\beta_{\mathrm{RFX}} = 2.8$, $P_{\mathrm{MCMC}} = 0.004$), with no effect of the irrelevant sensory feature. In scene trials only the main effect of scene is significant ($\beta_{\mathrm{RFX}} = 13.7$, $P_{\mathrm{MCMC}} < 0.001$). The standardized effect represents the expected value of the corresponding posterior beta estimate divided by its standard deviation. **f)** Computational modeling analysis based on the DDM reveals that tACS-induced behavioral modulations when motion is cued are specifically related to enhancing the rate of sensory evidence ($\beta_{\mathrm{RFX}} = 2.6$, $P_{\mathrm{MCMC}} = 0.0018$) while leaving all other parameters unaffected. **g)** A moving window analysis shows that the effect of the stimulation is online. The grey shaded area indicates the windows for which stimulation was turned on. Shaded areas around the lines indicate ± 1 SD of the posterior estimate of the interaction evidence*stimulation. The black bar at the top indicates $P < 0.05$ cluster corrected effects. **h)** We find that for motion-cued trials (left), out-of-phase stimulation significantly hinders performance ($P < 0.05$ cluster corrected). For scene-cued trials (right), in-phase stimulation improves performance marginally ($P < 0.05$ uncorrected).

of ongoing IFJ fluctuations induced by our stimulation protocol exclusively affect attention to the relevant (cued) sensory feature.

Dynamic evolution of IFJ top-down control modulations

The previous analyses were carried out during “stimulation on” periods, but do not allow interpreting: (i) whether these effects emerge exclusively during online stimulation; (ii) how they temporally evolve; and (iii) how these compare to periods without stimulation. To investigate this, we analyzed the temporal evolution of the in-phase vs out-of-phase stimulation effects (initially across sessions, Methods). When motion was cued, we found that the stimulation-induced attentional modulations emerged exclusively during the “stimulation on” periods and vanished immediately after the stimulation was switched off ($P < 0.05$ cluster corrected, Figure 2g), and in the correct direction but not significant when attention to scenes was cued. While these analyses reveal the robustness of the effects (when motion is cued and despite potential behavioral variability across sessions), these results do not allow us to conclude whether the stimulation-induced across-session modulations are driven by the in-phase, out-of-phase stimulation, or both. To investigate this, we analyzed the evolution of the stimulation effects within a single stimulation session relative to baseline periods of no stimulation (Methods). We found that out-of-phase stimulation robustly hindered discrimination performance

exclusively during "stimulation on" periods when motion was cued ($P < 0.05$ cluster corrected, Figure 2h), but this effect was not significant during in-phase stimulation ($P > 0.05$ cluster corrected, Figure 2h), and once again, these effects vanished immediately after the stimulation was switched off. Crucially, the interaction motion evidence*stimulation-condition was robustly significant in the hypothesized direction exclusively during the "stimulation on" periods ($P < 0.05$ cluster corrected, Figure 2h). When scenes were cued, in-phase stimulation significantly improved attentional performance only for a short time at the beginning of "simulation on" periods ($P < 0.05$ uncorrected, Figure 2h). Once again, these effects were not present for the distractor feature (Supplementary Figure 4.5). Thus, aligning periods of high excitability states in the IFJ with electric fields modulates non-spatial attentional behavior, with these effects being robust for motion perception.

Behavioral modulations of top-down control specifically affect sensory information processing

While our brain stimulation protocol appears to induce robust attentional influences in motion discrimination performance, these results do not clarify whether these behavioral modulations are indeed specific to boosting the perception of sensory evidence. We employed the drift-diffusion model (DDM), a well-established mathematical model of human choices that allows the possibility of disentangling how the manipulation of IFJ excitability states affects latent variables corresponding to distinct components of the decision process (Methods).

If it is true that IFJ excitability modulations specifically affect the degree of efficiency at which sensory areas accumulate sensory evidence, then we would expect in-phase stimulation to enhance the drift sensory rate latent variable. In line with our hypothesis, we found that in-phase stimulation during motion-cued trials boosted the rate of sensory evidence accumulation (interaction sensory evidence*stimulation condition $\beta_{\mathrm{RFX}} = 2.6$, $P_{\mathrm{MCMC}} = 0.0018$, Figure 2f), while leaving all other latent variables unaffected ($\beta_{\mathrm{RFX}} < 1.3$, $P_{\mathrm{MCMC}} > 0.09$).

Crucially, we investigated whether some of the above-mentioned differences in the modulation of top-down attentional control were related to our non-invasive brain stimulation intervention inducing oculomotor modulations. Analyses of eye tracking data show that there is no significant difference between the number of microsaccades in the different brain stimulation conditions (Supplementary Figure 4.2b). Together, our oculomotor and

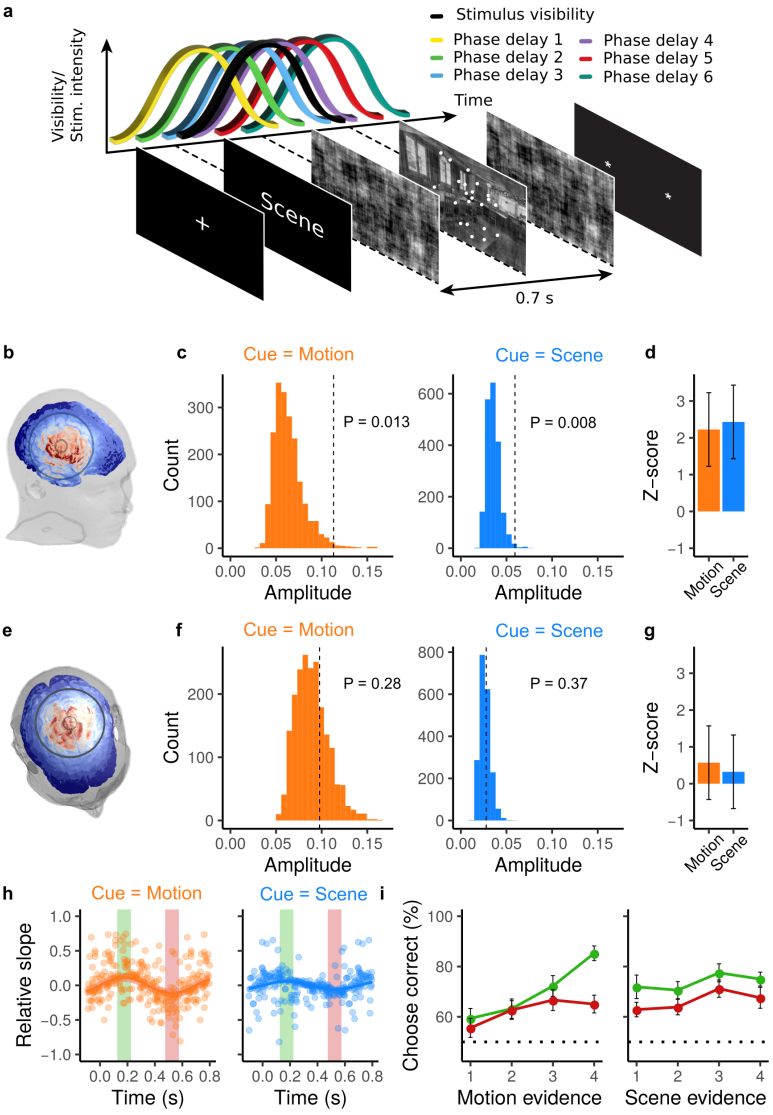


Figure 4.3. Phase-dependent influence of IFJ-tACS, but not Cz-tACS on non-spatial attention, Experiments 4 and 5. a) In this experiment, we introduced six stimulation delay conditions. The phase delays between the electrical and visual stimulation are evenly spaced out over one period of stimulation. We fit a sinusoidal function to the modulation of feature based attention as a function of phase delay, the amplitude of this function is the parameter of interest. **b)** In experiment 4 the centre of the electrodes is placed over the inferior frontal junction.

Figure 4.3: (continued) **c)** Since amplitude is a positive metric, we investigate its significance level by randomly shuffling all stimulation delay labels within participants and comparing the resulting distribution of estimated amplitudes with the estimated amplitude of the sinusoidal fit of the original data. We find that the amplitude of the fit of the original data is larger than 98.7% of the amplitudes of the generated distribution for motion trials and for scene trials this was 99.3%. **d)** The Z-scores of the empirical amplitudes as compared to the distribution of amplitudes expected to be found by chance is 2.2 for motion trials and 2.4 for scene trials. **e)** In experiment 5 the centre of the electrodes is placed on the location of the Cz electrode of the 10-20 EEG coordinate system. **f)** The control experiment shows that stimulating the motor cortex leads to no significant modulation of feature based attention to either motion or scenes (the empirical amplitudes are larger than 74% and 63% of the generated distribution of amplitudes, respectively). Therefore the effects of stimulation can not be contributed to the stimulation of an unrelated cortex or peripheral nerves. **g)** The Z-scores of the control experiment are 0.6 and 0.3 for respectively motion and scenes. **h)** Graph of the sinusoidal function of performance vs stimulation delay, with the estimated population-level parameters represented as a line, the shaded area indicates ± SD. The dots represent the individual data for each participant per stimulation delay condition after being aligned for individually estimated phase delays and intersects. The vertical green and red bars indicate the time windows of best and worst performance, respectively. **i)** Psychometric curves of the highest performance phase delay (green) and worst performance phase delay (red).

modeling analyses provide evidence that stimulation-induced attentional improvements are specifically related to boosting the degree of efficiency at which sensory areas accumulate sensory evidence [178, 179].

Non-invasive phase-dependent control of non-spatial attention

The next question we asked is whether the stimulation-induced attentional modulations necessarily require sensory tagging of the IFJ to rhythmic sensory manipulations. Additionally, we reasoned that the relatively weak effect for scenes in experiment 3 might be due to the multidimensional and non-local nature of the scene stimuli. In other words, because there is a larger activation area for scene recognition (Figure 1b) it might be harder to find the optimal timing of the stimulation if there is some degree of variability in the reaction of the cortical responses to sensory tagging across participants during the presentation of more complex sensory stimuli. To study these issues, we performed a new experiment (Experiment 4), where in each trial we presented a single stimulus that went in and out of "phase coherence" (Figure 3a, Methods). An additional feature in Experiment 4 is that we did not only stimulate in-phase or out-of-phase (as in Experiment 3) but tACS was applied at 6 different delays relative to the presentation of the sensory stimulus (Figure 3a). This allowed us to investigate whether non-spatial attentional modulations would fluctuate as a function of the phase of the tACS-induced

electric field. We found that the ongoing phase of the tACS signal induced significant modulations of behavioral performance when motion was cued ($z_A = 2.2$, $P = 0.013$, permutation tests, Figure 3c,d, Methods). This effect was smaller in overall effect-size terms but robustly significant when scene was cued ($z_A = 2.4$, $P = 0.008$ permutation tests, Figure 3c,d). The results of this experiment allow us to conclude that, first, continuous rhythmic sensory tagging is not necessary for inducing IFJ excitability modulations; second, non-spatial attention is related to excitability states of the IFJ which can be modulated as a function of exogenously applied electric fields.

Non-invasive phase-dependent modulations are likely not related to transcutaneous stimulation of peripheral nerves

Crucially, we conducted a new experiment (Experiment 5) to test whether the effects of tACS on non-spatial attention observed in Experiment 4 are (i) specific to the IFJ, (ii) are not due to our tailored design to induce a generalized oscillatory sensory tagging in the brain, (iii) are not due by transcutaneous stimulation of peripheral nerves [175, 180], and (iv) are not related to potential marginal influences of the electric field potentially reaching sensory areas. We identified (based on our neuroimaging data experiments) and stimulated a different brain structure to the IFJ that was in principle not related to non-spatial attention. The cortical area that we selected as the control target was the vertex (Cz location of the 10-20 EEG coordinate system, Figure 3e; a structure that is typically used as an active control site in non-invasive brain stimulation investigations studying higher cognitive functions [181]). All other experimental parameters were equal to those of experiment 4.

First, we confirmed that the electric fields in this active control condition do not greatly influence the IFJ, PPA, and V5. We found that the electric fields are virtually ineffective in these cortical areas (< 0.1 V/m for all voxels in the regions of interest, Supplementary Figure 4.4). Second, in the tACS-behavioral experiment, we found no significant modulations of behavioral performance as a function of the phase of the tACS-induced electric field for motion ($z_A = 0.6$, $P = 0.28$ permutation tests) nor for scenes ($z_A = 0.3$, $P = 0.37$ permutation tests). Thus, this active control experiment suggests that the modulatory effects of tACS on non-spatial attention observed in experiment 4 are indeed related to the stimulation of the IFJ and not due to the above-mentioned alternative explanations.

4.4. Discussion

Taken together, we developed a behavioral paradigm alongside a closed-loop non-invasive brain stimulation protocol that allowed us to predict and modulate with high temporal precision the IFJ excitability states during a non-spatial attention task. When the IFJ was predicted to be in a high excitability state, modulating it with tACS resulted in non-spatial attentional performance alterations. These effects were robust for motion evidence and replicated in a second experiment in which attentional modulations did not require a steady IFJ sensory tagging. While in general, the effects for scenes were in the hypothesized direction, they were not significant in the experiment with fixed in- and out-of-phase timings as identified for a different population sample in the EEG experiment. However, in experiment 4 we show significant modulation of attention to scenes, it could be that the variety of stimulation timings in experiment 4 makes it less sensitive to inter-individual differences. Given that sensory evidence for scenes is not exactly a unidimensional sensory feature that engages various features and a large portion of the ventral visual stream (contrary to motion perception), our stimulation protocol used in experiment 3 might require more specific sensory features to be more effective. Alternatively, future experiments aiming at modulating more complex sensory stimuli might profit from an individualized approach, for example by first performing an EEG experiment and estimating optimal timings of the stimulation per participant. As another option, based on the observation that most stimulation-induced effects in our study were in the hypothesized direction, it is tempting to speculate that increasing electric fields in the target area may result in more effective neural modulations and consequently more effective behavioral influences [174, 182]. In the control experiment we showed that the behavioral changes due to tACS are specific to the stimulation of the IFJ.

Our results show that the IFJ is causally involved in top-down non-spatial attention, which raises the question how the IFJ is connected anatomically and functionally. It has been shown with maps of probabilistic connectivities [183] that the IFJ has a high connection probability with both the fusiform face area (FFA) and the PPA, areas that are involved in high level non-spatial attention [51]. Furthermore, the authors show that the coherence between these areas increases in the tagging frequency as well as in high-gamma frequencies when attention to houses (PPA) and attention to faces (FFA) is exerted.

Temporal manipulations of sensory evidence in recent behavioral and neuroimaging studies led researchers to hypothesize that slow periodic neuronal excitability fluctuations in prefrontal structures shape the temporal dynamics of attention [167–170]. However, this hypothesis was questioned in a recent study suggesting that evidence for attentional rhythmic control is far from definitive due to statistical weaknesses in the analysis approaches [171]. While in our work we do not study the role of endogenous fluctuations but control them exogenously, our paradigm and results provide evidence that prefrontal excitability states are causally related to guide top-down attention. Here, we acknowledge that with our paradigm we cannot distinguish between modulating pure oscillatory neural activity or modulation of phasic activity on top of the exogenously controlled oscillation. However, irrespective of this consideration, our results support the theory that non-spatial attention relies on ongoing prefrontal excitability states which are likely regulated by slow oscillatory dynamics that guide goal-oriented behavior. Following up on our predator example, our findings entail that if the direction of the predator's movement behind the bush is the relevant feature, high excitability states in prefrontal structures regulating top-down attention would promote correct discrimination of the predator's direction of movement.

The methodologies developed in this work and the possibility of enhancing non-spatial attention may have important implications in disorders associated with the dysregulation of top-down control. For instance, lack of success in dietary behavior has been linked to reduced prefrontal top-down control of brain structures specialized in reward processing [63]. Failure to reduce the fear associated with traumatic experiences appears to be rooted in ineffective suppression of intrusive memories due to a lack of prefrontal top-down control over the hippocampus [184]. However, the brain-behavior relations in these examples remain purely correlative, and whether these functions depend on top-down control remains unknown. While the effects that we observed in our study appear to be effective during the stimulation periods, it has been recently shown that repeated application of tACS can have lasting beneficial effects [185]. The possibility of selectively modulating top-down control opens the door to understanding the mechanisms of attention in higher-level cognition, and to develop targeted therapies in disorders associated with top-down control dysregulation.

4.5. Methods

Participants

The study tested 142 healthy young volunteers: n = 20 participants took part in the fMRI study, Experiment 1 (mean age 25.6 years; range 21-36 years; 7 males); n = 23 in the EEG study, Experiment 2 (mean age 25.5 years; range 19-33 years; 7 males); n = 37 in the first tACS, Experiment 3 (mean age 25.8 years; range 18-40 years; 22 males); n = 37 in the second tACS study, Experiment 4 (mean age 24.3 years; range 18-35 years; 19 males); n = 25 in the third tACS study, Experiment 5 (mean age 25.1 years; range 19-36 years; 14 males). All participants had normal or corrected to-normal vision. Participants were instructed about all aspects of the experiment and gave written informed consent. None of the participants suffered from any neurological or psychological disorder or took medication that interfered with participation in our study. Participants received monetary compensation of 20 CHF/h for their participation in the experiment, in addition, they received 5 CHF/h if they got a mean performance score of 70% or higher. The experiments conformed to the Declaration of Helsinki and the experimental protocol was approved by the Ethics Committee of the Canton of Zurich.

Stimuli

To create a behavioral task in which it is necessary to employ non-spatial attention we created stimuli consisting of pictures and moving white dots spatially overlaid at the fovea. The visibility of these compound stimuli is dynamically modulated to follow a sinusoidal function, creating an opportunity for the visual cortices to entrain to the frequency of visual input. To make sure that the behavioral results are not contaminated by low-level confounds such as stimulus luminance or frequency spectra, we controlled the visibility of the stimuli using a phase-scrambling technique to preserve low-level image properties [186]. In brief, each image was Fourier-transformed, revealing pixel-by-pixel amplitude and phase information. A sequence of images is then generated by performing the inverse Fourier transform on a combination of the original amplitude spectrum with a modified phase spectrum. By changing the phase spectrum, we can control the recognizability of the image, while containing identical amplitude spectra and luminance to the original image. The phase consistency could range from 0.25 (almost no picture visibility) to 0.7 (the original picture is almost fully visible). The pictures represented either indoor or outdoor sceneries and were normalized to match mean luminance (SHINE toolbox, PsychToolbox). On top of the

pictures we presented 30 moving white dots, the direction of the average motion was either left or right. However, a percentage of the dots moved in a random direction; motion coherence ranged from 0.4 (almost no average direction) to 0.9 (clear average direction). The dots are shown in a circular aperture of 12°, centered at the fovea. Each dot covered $0.1° \times 0.1°$ of the visual angle and moved at 12° per second. The complete movie was sampled at the monitor's vertical refresh rate of 60 Hz. To synchronize the visual stimuli with the EEG recordings and tACS, two custom-built photosensitive triggers were placed on the sides of the monitor. This method was used in Experiment 2 to synchronize the EEG with the visual stimulation and in Experiments 3 and 4 to synchronize the visual stimulation with the electrical stimulation.

Behavioral paradigm

The behavioral paradigm is depicted in Figure 1a. During a trial participants first see a fixation cross, afterwards, we present a cue indicating to the participants whether they should pay attention to the motion or to the scene in the upcoming trial. Next, a sequence of four to seven compound stimuli (a scenery overlaid with moving dots) is presented. After the last compound stimulus disappears from the screen the participant should respond with a button press, only taking the last motion/scene into account. If the cue was scene the participant should press the left arrow key if the last scenery was indoor and right if it was outdoor. If the cue was motion the participant should press left for leftward motion and right for rightward motion. The participants have a maximum of 3 seconds to respond, if they fail to respond within this time the trial is automatically incorrect. Participants are instructed to be as fast and as accurate as possible. They were rewarded with an additional 5 CHF/hour for accuracies over 70%. Before starting the experiment the participants take part in a training session of 64 trials starting easy and increasing in difficulty level.

In the fMRI and EEG experiments, the first 64 trials consisted of the "no-attention" version of the task (Supplementary Figure 4.1). They were instructed to pay attention to the fixation cross and to press when the fixation cross changed orientation. The participants carried out this task with 86% and 89% accuracy for fMRI and EEG respectively, suggesting participant engagement in this task. The information presented on the screen was nearly identical to the information in the non-spatial attention task, with the exceptions of the words such as "Left" and "Right" which were replaced with nonsense text, and the fixation cross which was visible at all times and occasionally

rotated. This task was carried out before the participants were instructed about the non-spatial attention task, to avoid that they would pay attention to the visual stimulation other than the fixation cross (Supplementary Figure 4.1).

Eye-tracking measurements were acquired during all Experiments in this study to control for visual engagement during task performance (EyeLink 1000 Plus, SR Research, Ottawa, Ontario, Canada).

fMRI (Experiment 1)

fMRI acquisition

The fMRI data were acquired using a 3T Philips Ingenia with the visual stimuli being presented on an LCD monitor placed behind the participant. Participants looked at the stimuli using a mirror that was attached to the head-coil. Echo planar imaging (EPI)-blood oxygen level-dependent (BOLD) data were collected with a slice angle of 20° relative to the anterior–posterior commissure line, flip angle (FA) = 85°, echo time (TE) = 35 ms, repetition time (TR) = 2500 ms, 40 transversal slices (0 mm gap), and a 2.75 x 2.75 x 3.30 mm^3 voxel size (FOV = 222.75 x 222.75 x 128 mm^3). Subject-specific high definition structural T1 images were acquired through a magnetization-prepared rapid gradient echo (MPRAGE) sequence with the following parameters: FA = 8°, TE = 3.6 ms, TR = 7.7 ms and a 1 x 1 x 1 mm^3 voxel size (FOV = 240 x 240 x 160 mm^3).

fMRI analyses

Analysis and pre-processing of the data was performed in FSL's Analysis Tool FEAT v6.0.0, this included a BET brain extraction, slice timing correction, motion correction using MCFLIRT, a Gaussian spatial smoothing with a full width at half maximum (FWHM) of 5 mm, and a high pass temporal filtering with a cut-off of 100 s. Images were then spatially normalized using FLIRT (FMRIB's Linear Image Registration Tool) registering the low-resolution functional images to the high-resolution structural image and then using FNIRT (FMRIB's Nonlinear Image Registration Tool) the images were warped onto the reference brain in the Montreal Neurological Institute (MNI) coordinate space.

First level analysis was performed with FILM (FMRIB's Improved Linear Model) based on general linear modelling (GLM) with the canonical hemodynamic response function (HRF) as its base function. Explanatory variables

included in the analysis of the attention task performance are attention to scene, attention to motion, response to scene, and response to motion. A contrast was defined for attention to scene vs attention to motion. For the passive viewing analysis, the explanatory variables were visual stimulus presentation and button presses. Group-level analysis was performed using FLAME (FMRIB's Local Analysis of Mixed Effects Tool). Contrasts were defined for attention to scene vs visual stimulus presentation and attention to motion vs visual stimulus presentation. Z-statistic images were thresholded at $Z > 2.6$ and a cluster correction was applied at a threshold of $P < 0.05$.

EEG (Experiment 2)

EEG acquisition and preprocessing

EEG was acquired at 500 Hz using a high-density net (128 Channels Geodesic Sensor Net, Magstim EGI, Eugene, USA). EEG data preprocessing and analysis were performed using the Fieldtrip toolbox ([187], Donders Institute for Brain, Cognition and Behaviour, Radboud University, the Netherlands) in MATLAB (R2019b, MathWorks inc., Natick, USA). Line noise was removed using a discrete Fourier transform filter. The data were re-referenced to a common average reference and epoched into 0 to 2.8 s trials to include the first four tagging cycles of each trial. Bad channels and noisy trials were removed based on visual inspection.

To quantify the neural entrainment to the visual stimulation, we computed the debiased weighted phase lag index (dWPLI) [188] at 1.43 Hz between the sensor data and the imposed visibility sine wave tagging with a frequency of 1.43 Hz. We used the dWPLI as it is a phase-synchronization index that is robust to sample-size bias and spurious connectivity driven by volume conduction. This computation was performed separately for the attention and no-attention tasks and the comparison is shown in Figure 1f.

To localize which neural structures were entrained by the visual stimulation, source reconstruction was performed using linearly constrained minimum variance beamforming [189]. This analysis estimates the time series in each dimension for each voxel in the brain by computing spatial filters based on the location of the sensors. To reduce the dimensionality, single value decomposition was used to compute the projection with the largest variance for each voxel. To quantify the entrainment to the visual stimulation, a similar approach to the sensor-level analysis was used. The dWPLI at 1.43 Hz was

computed for each voxel between the time series of the projection with the largest variance and an artificial signal of a sine wave with a frequency of 1.43 Hz corresponding to the visual stimulation. The source reconstruction and dWPLI computation were performed for data from the attention and no-attention tasks separately. We identified for each subject the voxel with the highest dWPLI near the IFJ and near the visual cortex in the attention task. The time series of these voxels for one example participant is shown in Figure 1e. To compute the delay between the visual stimulation and the neural oscillations, the source time series were band-pass filtered using a FIR filter with a cut-off of 1.43 ± 0.01 Hz. We then measured the latency of the third peak of the time series and subtracted 700 ms*2.5 = 1.750 ms which corresponds to the third peak of the visual stimulation. These delays are presented in Figure 1j and were used to determine the timing of the electrical stimulation in Experiment 3 (Supplementary Figure 4.3). Prior to source analysis, the data were low-pass filtered at 20 Hz using a two-pass hamming filter. The data were in addition high-pass filtered at 0.3 Hz with a Butterworth filter for the visualization in Figure 1e. In this Experiment, 4 participants were excluded due to excessive noise in the EEG recordings.

EEG statistical analyses

The sensor dWPLI values were compared between the attention and no-attention tasks. Cluster correction was performed by generating Markov chain Monte Carlo simulations with 5,000 permutations to determine the multiple comparison cluster correction at $P < 0.05$ based on the null distribution of clusters thresholded at $P < 0.01$.

A similar approach was used for the source-level statistics. The computed dWPLIs for each voxel were compared between the attention and no-attention tasks. Cluster correction was performed by generating Markov chain Monte Carlo simulations with 5,000 permutations to determine the $P < 0.05$ threshold of the null distribution of clusters of voxels thresholded at $P < 0.01$.

Brain stimulation (Experiments 3-5)

tACS application

For the application of tACS, we used a current stimulator (DC-stimulator, neuroConn) with a manual stimulation protocol controlled by MATLAB. We used two concentric ring electrodes (active electrode diameter 2 cm,

return electrode inner diameter 7.5 cm, outer diameter 10 cm). Following the visual sensory tagging of our behavioral paradigm, tACS was applied at a frequency of 1.43 Hz (period 0.7 s). The amplitude was determined for every participant individually with custom-written code. The maximum current used was 4 mA peak-to-peak. At the beginning and at the end of each stimulation block the current was ramped up and down over the first and last 10 s, respectively. As a baseline condition, we applied sham stimulation, for which we ramped up the current to its maximum amplitude over 10 s, before turning it off again. The tACS was applied continuously during the stimulation block and precisely synchronized with the visual stimuli using two photosensitive triggers attached to the monitor and custom-written code in MATLAB which was synchronized with the computer controlling the visual input and behavioral output of our participants.

The concentric ring electrodes are placed on the scalp of the participant with the centers over the left and right IFJ. The location on the scalp that is nearest to the left and right IFJ was estimated for test participants using a structural T1 MRI scan and neuronavigation. The average IFJ location converged in all cases into channels 117 and 128 of a EGI Geodesic 128-channel EEG cap, which was used to find the IFJ location in preparation for the tACS experiments for all participants. A topical anesthetic (EMLA cream 5%) is used to numb the skin under the active electrodes. This procedure reduces the skin sensations induced by transcranial stimulation which makes the stimulation more comfortable.

Electric field predictions

To investigate the strength of tACS exposure during the experiment, an in silico model was developed. Electromagnetic (EM) simulations were executed to predict electric field (E-field) exposures within the brain and the target region, namely the IFJ. The EM simulations were performed using the Sim4Life (ZMT Zurich Med Tech AG, Zurich, Switzerland) platform for computational life-science investigations, using the detailed anatomical MIDA head model [190]. The model distinguishes 37 tissue classes, of which the electric conductivities were assigned according to the IT'IS Low Frequency Database V4.1 [191]. The analysis pipeline consisted of the following steps: (i) the creation of electrode models and their placement on the skin of the MIDA (Virtual Population, IT'IS Foundation, Zurich) head model; (ii) identification of the anatomical target region and positioning of the electrodes;

(iii) execution of the EM simulations; and (iv) estimation of the predicted E-field distributions within the IFJ and the rest of the brain.

The target region (IFJ) in the MIDA model was identified by registering the MIDA's brain T1 images with the open-access Brainnetome Atlas [192] using FSL v5.0 FLIRT, by importing the transformed atlas in Sim4Life, and aligning it with the MIDA model. The atlas defines 246 brain areas, including left and right IFJ, which were applied as masks to the MIDA model (Supplementary Figure 4.4a, b). While the stimulation target and positioning of electrodes were selected during a brain mapping procedure and defined in MNI space, in addition to coregistration of the MIDA with the Brainnetome atlas, we developed a pipeline aimed at identifying the MNI coordinates in the MIDA brain. For this pipeline, first, the MIDA brain mask was normalized to MNI space, in which the IFJ area (MNI left IFJ = -54, 12, 34 mm, MNI right IFJ = 54, 12, 34 mm) and electrode coordinates (MNI left electrode = -60, 12, 38 mm, MNI right electrode = 60, 12, 38 mm) were identified. After that, 14 x 14 x 14 mm masks were drawn around the locations of the targets and electrodes. Finally, the normalized MIDA brain together with the new masks was coregistered with the initial MIDA brain and imported into Sim4Life. At the end of this procedure, we compared the location of the target defined in the MNI space and the IFJ determined with the Brainnetome atlas, and concluded that these targets have the same positioning in the MIDA brain (Supplementary Figure 4.4b). This pipeline was implemented with the SPM12 toolbox in MATLAB R2019a.

The electrode geometries were created in Sim4Life using the constructive geometry functionality in Sim4Life. Two cylindrical electrodes were created with radius = 1 cm and were placed above the left and right IFJ, with two surrounding ring electrodes (inner radius = 4 cm, outer = 5 cm, Supplementary Figure 4.4a). Two sensor boxes were placed around the central electrodes to evaluate the current and normalize the E-field distribution to the total current.

EM simulations were executed using Sim4Life's rectilinear version of the 'Electro Ohmic Quasi-Static' finite element method (FEM) solver [193]. The model geometry was discretized with a grid resolution between 0.5 and 0.75 mm — identified through a convergence analysis — with the highest refinement near the electrodes. An EM simulation was executed for each electrode, by assigning Dirichlet (voltage) boundary conditions of +1V to the central electrode and -1V to the ring electrode, while assigning the other electrodes to Perfect Electric Conductor (PEC). The total E-field was

calculated using the superposition principle considering that the two currents are in-phase (i.e., same frequency), and the focality and intensity of the stimulation were extracted on the target region.

Additionally, we ran a simulation for experiment 5 placing the centre electrode on the Cz location of the MIDA model surrounded with the ring electrode, using the same computational parameters as for the previous simulations. The results of this model demonstrate that the E-field within the target and control regions was minimal and could not lead to activation of these areas even under 4 mA peak-to-peak stimulation: right and left IFJ (mean= 0.04, sd = 0.01, $P_{99} = 0.07$), PPA (mean= 0.008, sd = 0.002, $P_{99} = 0.01$) and MT/V5 (mean= 0.007, sd = 0.001, $P_{99} = 0.01$; Supplementary Figure 4.4e,f).

Experiment 3: In-phase vs out-of-phase stimulation

In Experiment 3, two stimulation conditions were used, in-phase and out-of-phase stimulation. During in-phase stimulation, the visibility of the stimulus precedes the voltage over the electrodes by 95 ± 18 ms, see Supplementary Figure 4.3. During out-of-phase stimulation timing of the voltage over the electrodes is shifted by 180° (350 ms).

The tACS experiment consists of two sessions, each with 320 trials. In each session, either all motion trials are stimulated in-phase and all scene trials are stimulated out-of-phase or vice versa. For half of the participants, pseudo-randomly chosen, the first session consisted of the in-phase stimulation condition for trials in which scene is cued, and out-of-phase stimulation for trials in which motion is cued. For these participants, the second session consisted of in-phase stimulation for motion trials and out-of-phase stimulation for scene trials. For the second half of the participants, the order of the stimulation conditions was reversed. A single session is divided into four blocks. Only in blocks two and three stimulation is turned on. Blocks one and four consisted of sham stimulation. In half of the trials of blocks one and four, the participant is cued after the visual stimulus has disappeared, thus making the participant pay attention to both the scene and motion during stimulus presentation. Participants with discrimination performance $< 55\%$ in block one (suggesting nearly random choice selection and therefore poor engagement) were excluded from the data analyses, resulting in the exclusion of 4 participants.

Experiment 4: tACS phase-dependent effects

In the phase-dependent effects tACS experiment, we make use of six stimulation delays spaced out evenly over the period of a single visual stimulus period. To maximize statistical power over the 6 different conditions (which are in turn divided into the 4 sensory evidence levels used in Experiment 3), the experimental session consisted of the continuous application of tACS. The stimulation delays were pseudo-randomly assigned on a trial-to-trial basis. Thus, this experiment allows us to study the relationship between the ongoing phase of the tACS stimulation relative to the presentation of the visual stimulus. Details regarding the statistical analyses are described in the next section. Participants with discrimination performance $< 55\%$ in block one (suggesting nearly random choice selection and therefore poor engagement) were excluded from the data analyses, resulting in the exclusion of 5 participants.

Experiment 5: control stimulation location

We placed the center of the electrodes over the Cz location of the 10-20 EEG coordinate system, therefore stimulating the motor cortex. All other experimental parameters were equal to those of experiment 4. Participants with discrimination performance $< 55\%$ in block one (suggesting nearly random choice selection and therefore poor engagement) were excluded from the data analyses, resulting in the exclusion of 3 participants.

Eye tracking

Eyetracking (EyeLink 1000 Plus) was used to check the participants' eye movement during stimulus presentation. A chinrest was used to keep the distance between the participants and the screen constant (55 cm). Extraction of microsaccade data was analyzed using the widely adopted approach described by Engbert and Kleigl [194]. We focused on the combination of saccades and microsaccades (saccades $<$ 1 degree of visual angle) occurring within the first four tagging cycles of each trial (the first 2.8 seconds of stimulus presentation).

Behavioral analysis and statistics

Mixed-effects model of sensory discrimination behavior

A logistic mixed model was implemented to investigate the effect of stimulation (in-phase or out-of-phase) on the participant's sensory discrimination as a

function of both the cued and the distractor sensory evidence. Trials in which the participant is cued to pay attention to motion, the motion evidence is the main explanatory variable, while scene evidence is a distractor that should be ignored and vice versa. The log-odds for making the left or right decision is given by

$$\begin{aligned}\bar{\beta} = \beta_0 + \beta_1 * \text{Motion} + \beta_2 * \text{Scene} + \beta_3 * \text{Stim} \\ + \beta_4 * \text{Motion} * \text{Stim} + \beta_5 * \text{Scene} * \text{Stim},\end{aligned} \tag{4.1}$$

where the probability of selecting "right" and explaining the participant's (s) response $y_{i,s} \in 0, 1$ (with $y = 0$: left, $y = 1$: right) in trial i is given by

$$\begin{aligned}\theta_{i,s} &= 1/\left(1 + e^{-\bar{\beta}}\right) \\ y_{i,s} &\sim \text{Bernoulli}(\theta_{i,s}).\end{aligned} \tag{4.2}$$

A positive interaction effect between stimulation and sensory evidence (motion or scene) indicates that the corresponding sensory information influences participant's behavior more strongly in the in-phase stimulation condition compared to out-of-phase.

Dynamic evolution analyses of stimulation effects

To study how the stimulation influenced task performance over time, a moving window analysis of tACS influences on behavior was performed with a window length of 90 trials. For each window, a logistic mixed-effects model similar to the one described above was fitted to the behavioral choice data. In the corresponding figures, we report the standardized interaction evidence*stimulation with the error denoting the ± values of 1 SD. The interaction effects were cluster corrected at $P < 0.05$ by constructing a null distribution of cluster sizes, based on shuffling the labels of the stimulation phase data within participants.

Computational model

Our brain stimulation protocol appears to induce attentional influences in sensory discrimination performance. However, these results do not clarify whether these behavioral modulations are indeed specific to boosting the perception of sensory evidence. A way to clarify this would be to combine tACS during neuroimaging, however, due to technical and safety aspects, we were not able to apply current intensities above 2 mA peak-to-peak, while

in our behavioral studies we applied currents of up to 4 mA peak-to-peak. Therefore it was not possible to combine tACS and fMRI with the protocol developed here. Nevertheless, this question can be tackled with the use of computational models.

We analyzed the influence of tACS on the discriminability of the cued sensory feature with a prominent mathematical model of two-alternative decisions, the drift-diffusion model (DDM), which incorporates both observed choices and reaction times (RT) to decompose the decision process into distinct latent variables corresponding to distinct aspects of the choice process: (i) the efficiency of sensory evidence accumulation, known as the drift rate (δ); (ii) any bias in the choice process (β); (iii) the amount of evidence required to make a decision, known as the decision threshold (α); and (iv) the delay in the onset of evidence accumulation, the non-decision time (τ).

The decision-making model implemented here is based on a simple one-dimensional Wiener process: a dynamical system where the state of evidence $X(t)$ at time t evolves via the stochastic equation $\frac{\mathrm{d}X(t)}{\mathrm{d}t} \sim \text{Normal}\left(\delta, \sigma^2\right)$ where δ represents the quality of information processing defined as $\delta = kE$, where E represents the sensory evidence level (i.e., the stimulus visibility in our task) and k a variable that linearly scales the evidence. For initial conditions, where β represents an initial bias in the process, it is assumed that the system makes a decision ζ (left or right) at time t whenever $X(t) >= \alpha$ (right) or $X(t) < 0$ (left). In addition, we accounted for visual processing and corticomuscular response delays via the non-decision time parameter τ (the RT in each trial is defined as RT$= t + \tau$). The goal is to find the Wiener distribution, Wiener$(\delta, \alpha, \tau, \beta)$, that best explains the distribution of empirical choices $y(\zeta, \text{RT})$. To this end, we implement a hierarchical Bayesian model where each individual data point $y_{i,s}(\zeta, \text{RT})$ follows a Wiener distribution

$$y_{i,s} \sim \text{Wiener}(\delta, \alpha, \tau, \beta), \tag{4.3}$$

with indices s for subjects ($s = 1, \ldots, N_{\text{subjects}}$) and i for trials ($i = 1, \ldots, N_{\text{trials}}$).

Given that in our study we use a hierarchical Bayesian data analysis framework, this allows the convenient possibility of studying the effects of a given tACS stimulation condition (e.g., in-phase stimulation) on a latent variable during a baseline condition (e.g., the drift-rate modulator k during out-of-

phase stimulation or baseline trials). Thus, we study the (potential) relative change of a given latent variable $\theta \in \{k, \alpha, \tau, \beta\}$ as follows

$$\theta_{s,i} = \theta_{\text{base},s} + \beta_s^{\theta} * D_i, \tag{4.4}$$

where $D \in \{1, 0\}$ denotes whether the modulator condition (e.g., in-phase stimulation in our example) was present ($D = 1$) or not ($D = 0$) in each trial i. The subscript s denotes that the effect is participant-specific which is modeled as a random-effects factor under the assumption that it is drawn from population distributions $\theta_{\text{base},s} \sim N(\theta_{\text{base}}, \sigma_{\text{base}})$ and $\beta_s^{\theta} \sim N(\beta^{\theta}, \sigma^{\theta})$ where $\theta_{\text{base}}, \beta^{\theta}$ and $\sigma_{\text{base}}, \sigma^{\theta}$ determine the mean and the standard deviation of the population distributions, respectively.

Sinusoidal model (Experiments 4 and 5)

The aim of Experiments 4 and 5 is to study whether the ongoing tACS phase relative to a single stimulus presentation modulates non-spatial attention behavior. We synchronized the ongoing tACS peak at six equally spaced phase delays over the period of one full stimulation period (Figure 3a). To study the influence of the delays in a parsimonious parametric model, first, we performed a separate logistic regression for each participant and each stimulation delay condition as follows:

$$\begin{aligned} \theta_i &= 1/\left(1 + e^{-(\beta_{s,d} + E_i * \delta_{s,d})}\right) \\ y_i &\sim \text{Bernoulli}(\theta_i), \end{aligned} \tag{4.5}$$

where $y_i \in \{0, 1\}$ denotes the trial-to-trial choices in each trial i as a function of E_i, which denotes the amount of motion evidence in the trials in which motion was cued and the amount of scene evidence in those trials scene was cued. $\beta_{s,d}$ is a subject (s) and stimulation delay (d) specific bias term, and $\delta_{s,d}$ corresponds to a subject and stimulation delay specific slope. Next, we fit a sinusoidal function through the slope parameters of the logistic regression as a function stimulation delay:

$$\begin{aligned} \mu_{s,d}(\tau_d) &= \beta_s + A_s \sin\left(\frac{2\pi\tau_{s,d}}{6} + \phi_s\right) \\ \delta_{s,d} &\sim N(\mu_{s,d}, \sigma_d), \end{aligned} \tag{4.6}$$

where $\delta_{s,d}$ is the population distribution of the subject-specific psychometric slopes for each stimulation delay (d) obtained in Eq. 4.5, and $\tau_{s,d}$ the timing

of the different tACS phase delays. β_s represents the subject (s) specific offset of the sinusoidal function with amplitude A_s. Parameter ϕ_s determines the phase shift which was parameterized as a von Misses distribution

$$\phi_s \sim \frac{\exp(\kappa \cos(x - \phi))}{2\pi I_0(\kappa)}, \tag{4.7}$$

initialized with a flat prior, that is $\kappa = 0$, where I_0 is the modified Bessel function of the first kind of order 0.

Here it is important to emphasize that the key parameter determining a tACS phase-delay modulation is the population level estimate of the sinusoidal amplitude, which is estimated departing from an exponential prior distribution

$$A_s \sim \lambda e^{-\lambda A} > 0, \tag{4.8}$$

with a conservative prior by setting $\lambda = 4$. However, we found that our results are largely insensitive to the selection of this prior. Please note that this conservative prior promotes smaller amplitudes, as psychometric slopes larger than 1 are unlikely. Given that this parameter is by definition positive, the significance of the expected amplitude at the population level $\mathbb{E}[A]$ was determined by comparing this value to a null distribution of amplitude expected values $\mathbb{E}[A]_{\text{rand}}$, based on shuffling the labels of the stimulation phase data within participants and repeating 5,000 times the procedure described in Eqs. 6-8 to estimate each $\mathbb{E}[A]_{\text{rand}}$. To compare the effects of tACS across conditions, we obtained a standardized estimate of the amplitude modulation effect z_A. Assuming that the null distribution of amplitude expected values $\mathbb{E}[A]_{\text{rand}}$ approximates a normal distribution, we define the standardized estimate of the amplitude modulation effect z_A as

$$z_A = \sqrt{2}\,\text{erf}^{-1}(2P - 1), \tag{4.9}$$

where P is the proportion of samples of the null distribution smaller than $\mathbb{E}[A]$, and $\text{erf}^{-1}(x)$ is the inverse of the error function $\text{erf}(x)$

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{4.10}$$

Statistical inference

All mixed-effects models in this study had varying subject-specific latent variables unless otherwise specified in each model description. Posterior inference of the parameters in the hierarchical models was performed via the

Gibbs sampler using the Markov Chain Monte Carlo (MCMC) technique implemented in JAGS, assuming flat priors for all population-level parameters (unless otherwise specified). For each model, a total of 100,000 samples were drawn from an initial burn-in step and subsequently, a total of new 100,000 samples were drawn with three chains (each chain was derived based on a different random number generator engine, and each with a different seed). We applied a thinning of 100 to this final sample, thus resulting in a final set of 1,000 samples for each parameter. We conducted Gelman–Rubin tests for each parameter to confirm the convergence of the chains. All latent variables in our Bayesian models had $\hat{R} < 1.05$, which suggests that all three chains converged to a target posterior distribution. We checked via visual inspection that the posterior population-level distributions of the final MCMC chains converged to our assumed parametrizations. For all random effects reported here, the reported value corresponds to the mean of the standardized posterior distribution, and the "P-values" reported for these regressions are not frequentist P-values but instead directly quantify the probability of the reported effect differing from zero (P_{MCMC}). They were computed using the posterior population distributions estimated for each parameter and represent the portion of the density functions that lies above/below 0 (depending on the direction of the effect). The standardized effects of the hierarchical mixed-effects models reported in the main text were obtained by dividing the expected value of the corresponding posterior beta estimate by its standard deviation.

Data and code Availability

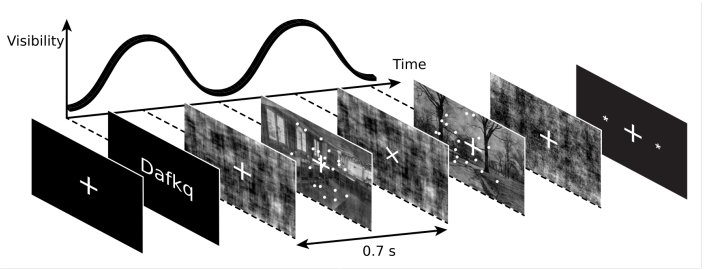
Data and code will be made openly available upon manuscript acceptance.

Author Contributions

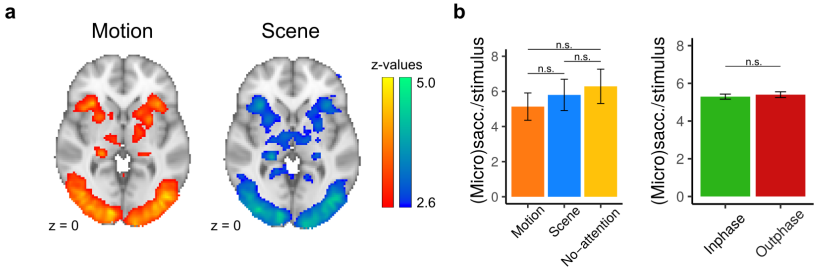
J.B. wrote experimental code, collected fMRI and tACS data, performed data analysis, and implemented computational models. J.H. wrote experimental code and performed EEG data analysis. V.B., A.C., and E.N. performed computational modeling of electric fields. F.G. collected EEG data, performed EEG analysis, and troubleshot experimental setups. J.B., J.H., V.B., M.G., and R.P. contributed to the conceptual development of the experimental paradigm, discussed the results, and wrote the paper. R.P. raised funding.

Acknowledgements

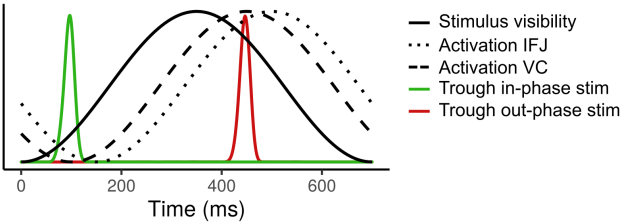
This work was supported by a European Research Council (ERC) starting grant (ENTRAINER) to R.P., an ETH Grant (ETH-25 18–2) to R.P., and a ZNZ PhD grant to J.B. and R.P. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 758604). We would also like to thank Jin Qiuhan for helping pilot the experimental paradigm, Irmak Gezginer for helping collect the EEG data, Margaux Quiniou, Hilde van der Pol, and Sebastian Warma for helping collect the tACS data, and Marijke Compaijen for proof-reading the manuscript.



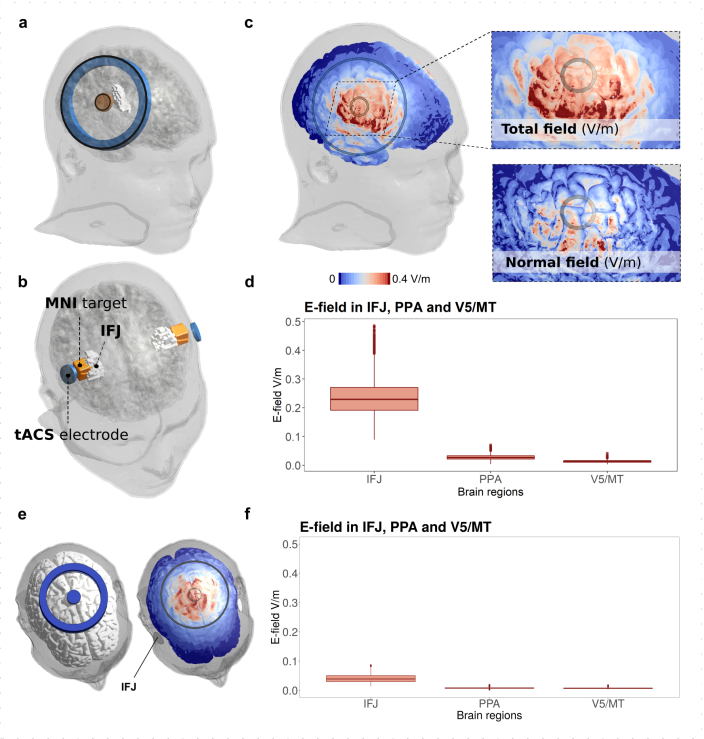
Supplementary Figure 4.1. Control "no-attention" task in Experiments 1 and 2. To study brain activity purely evoked by the visual input, without the influence of attention, in the fMRI and EEG experiments participants first carried out a version of the task without non-spatial attention. They were instructed to pay attention to the fixation cross and to press a button when the fixation cross made a 45° orientation shift. These orientation shifts would happen at random intervals, uniformly distributed between 5 and 30 seconds. The visual information on the screen was identical to the non-spatial attention task, with the exception that all text was replaced by nonsense text and the fixation cross was visible at all times.



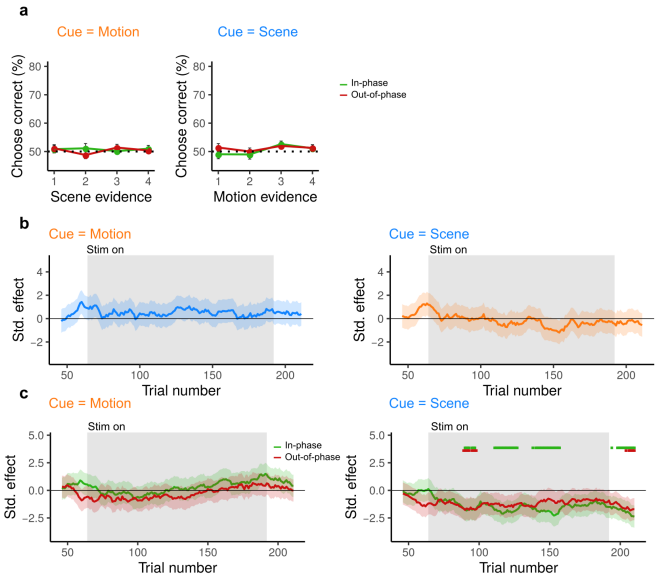
Supplementary Figure 4.2. Wide activations of the visual cortex and microsaccade analysis. a) Using fMRI without contrasting attentive vs. unattentive states we find a wide activation of the visual cortex, both when participants are cued for motion and for scenes. **b)** Analysis of the eye tracking data shows no significant differences in the number of microsaccades per stimulus presentation in which motion or scenes were cued, or in unattentive states. There are also no significant differences in microsaccades per stimulus for in-phase or out-of-phase stimulated trials (All pairwise combinations of paired t-tests $P > 0.4$).



Supplementary Figure 4.3. Visual stimulus and transcranial electrical stimulation timing. Based on the results obtained in our EEG experiment, we found that the visual cortex (VC) and the IFJ get entrained to the visual stimulation. The delay between visual stimulation and the response in the visual cortex was about 100 ms and 150 ms for the IFJ. Using photosensitive triggers on the monitor we could record the exact timing of the visual stimulation and compare it to the ongoing electrical stimulation. The timing of the trough of in-phase stimulation (mean = 95 ms after the trough of visual stimulation, SD = 9) is represented in green and out-of-phase (mean = 445 ms after the trough of visual stimulation, SD = 15) in red. By timing the tACS waveform a few ms before the tagged slow rhythmic fluctuations in visual and prefrontal areas, we hypothesized that we could maximize the influence of the stimulation on behavior.



Supplementary Figure 4.4. tACS electric field predictions and coregistration procedure. a) Model geometry with the identified IFJ (in white) and the stimulation electrodes (round in yellow and ring in blue) above the target. The contour of the electrodes is highlighted in black. The identification of the IFJ in the MIDA was done by coregistering the MIDA with the Brainnetome atlas. **b)** The IFJ segmented from the atlas (in white) and target electrode (in blue) displayed together with the cortical target (in yellow) identified in the MNI space. **c)** Left: absolute E-field distribution on the cortical surface for an input current of 4 mA (peak-to-peak). Right: Surface views of the total E-field magnitude (top view) and of the normal E-field component to the cortex (bottom view) which is considered to be the principally relevant E-field component coupling with the electrophysiology of pyramidal neurons. **d)** Boxplots represent the distribution of the E-field in each voxel within the IFJ (0.24 ± 0.06), PPA (0.028 ± 0.008), and MT+ (0.013 ± 0.004). These results indicate that the relevant sensory areas in this study were not affected by the application of our tACS protocol. Moreover, the fact that influences of our tACS protocol on non-spatial attention were larger in motion relative to scenes cannot be explained by differences in E-field strength as these are negligible in both sensory areas. **e)** In the control experiment (Experiment 5) the electrodes are placed on the Cz location of the 10-20 EEG coordinate system, therefore stimulating the motor cortex. **f)** Same as in **d** but for the electrical fields as produced in the control experiment. The distributions of the E-field is within the IFJ (0.04 ± 0.01), PPA (0.008 ± 0.002), MT+ (0.007 ± 0.002).



Supplementary Figure 4.5. The tACS-induced behavioral changes in sensory discrimination of the cued feature are not likely to be influenced by behaviorally-induced effects in the irrelevant feature. a Participants are not distracted by uncued evidence irrespective of the stimulation phase. When participants are cued to motion, scene evidence does not influence their decisions and when cued to scenes motion evidence does not influence their decisions. **b** A moving window analysis shows that there is no significant effect of the stimulation on distraction through time. The grey shaded area indicates the windows for which stimulation was turned on. Shaded areas around the lines indicate $\pm$ 1 SD of the posterior estimate of the interaction evidence*stimulation. **c** Comparing in-phase and out-of-phase stimulation against baseline performance we find that for trials in which motion is cued neither in-phase nor out-of-phase distraction levels differ from baseline ($P < 0.05$ uncorrected). For scene trials at certain time point participants are significantly distracted by motion evidence. The negative sign of this effect indicates that participants tend to choose outdoor (right button) when motion evidence is towards the left and indoor (left button) when motion evidence is towards the right.

5

SNACK CHOICES BUT NOT WILLINGNESS TO EAT RATINGS PREDICT BMI GROUP

J. A. Heng, C. Joray, D. Popelka, D. Herzig, L. Bally, Rafael Polanía. Snack choices but not willingness to eat ratings predict BMI group. In preparation.

Contributions

Conceptualization, Experimental software design, Data collection, Data analysis, Writing

5.1. Abstract

There is growing evidence that obesity is associated with alterations in dietary decision-making. However, a comprehensive description of these alterations is lacking. We develop models of dietary decision-making based on the idea that the value of food items depends on the sum of their nutritional and non-nutritional attributes. We find that participants are influenced by many nutritional and non-nutritional attributes when asked to rate their willingess to eat a snack or to make a choice between two snacks. We find that participants are influenced by their beliefs about the nutritional attributes rather than the true attribute values. We also repeat previous findings that overt attention has an influence on choice. We compared the behavior of participants with and without obesity, and found they nutritional and non-nutritional attributes had similar influence on their willingess to eat ratings. However, we found differences in the influence of these attributes and overt attention on the choices of the participants. This allowed our models to classify with an AUC-ROC of 0.8 the BMI group of individuals based on their choices. Importantly, these models are easy to interpret, and can indicate which choices and which attributes carry information about BMI group.

5.2. Introduction

What makes a food item valuable? A possible response is that the value of a food item stems from its nutritional attributes. For instance, someone who likes sugar and fat may prefer to eat a piece of cake over a carrot. Healthiness can also be considered a nutritional attribute and may also play a role in the valuation process. Our decision maker choosing between a piece of cake and a carrot may weight healthiness more than sugar and fat, and reconsider their choice. In addition to the nutritional attributes, food items may be given value based on their non-nutritional attributes. For instance, consider two cakes made with the same ingredients. If one of these cakes is well decorated, you will likely prefer to eat it over a cake that is not. Therefore, we can consider that food valuation is a process that depends on the weighing of different nutritional and non-nutritional attributes.

Previous studies have investigated how nutritional attributes influence value. These studies typically correlate nutritional attributes of food items with the willingness to pay for these items. The results point to the influence of calories, [195], fat, carbohydrate, sugar, protein, sodium, vitamins [196] and the combination of fat and carbohydrate [197]. Some studies investigate if the food item valuation is rather influenced by the participants' belief about the content of nutritional attributes or their true content, with mixed results [195, 196]. There has also been a focus on studying the different effects of healthiness and tastiness ratings on food choice [63, 198, 199]. However, with the exception of the price of the item, these studies typically do not consider non-nutritional attributes.

In comparison to nutritional attributes, the influence of non-nutritional attributes on the valuation process has been less studied. This gap is unjustified, as non-nutritional attributes such as price [200], color [201] and texture [202] can influence taste perception. Lee and Hare [203] found that in addition to healthiness and tastiness, appearance and texture influence valuation. Consumer behavior research has been interested in the influence of packaging of the items, and has found that packaging shape and color influence the consumers willingness to pay for an item [204]. Altogether, the valuation of food items depends on nutritional and non-nutritional attributes.

Another line of research has considered the role of attention in decision-making. When faced with a choice, our brains may not be able to process all options equally. It has been suggested that the gaze dynamics (i.e., which items the participant looks at), which can be measured by eye-tracking, can

reveal where the participant pays attention and therefore which options are considered at a given time. Following this approach, it has been found that participants tend to choose more the items that they look at longer [64, 65]. This effect has been found for non-dietary goods, but also for food items. Considering these attentional mechanisms may help to understand dietary decision-making.

Understanding dietary decision-making is relevant to understand disorders in which dietary decision-making is altered, for example obesity. It has become clear that dietary decision-making is altered in obesity [60]. However, it is still unclear which decision-making processes are altered. Several studies have investigated these alterations. For example, no difference has been found between liking and wanting milkshakes between individuals with and without obesity [205]. However, the effect of the combination of fat and carbohydrate on willingness to pay for a food item has been observed in healthy participants but not in individuals with obesity [206]. Numerous studies have investigated differences in attention to food items in obesity [207–213]. A meta-analysis concluded that individuals with obesity had no attentional bias to food images as measured by eye-tracking metrics [214]. However, to the best of our knowledge, there has been no study investigating the role of attention in decision-making in obesity.

The goals of this study are to investigate the influence of nutritional and non-nutritional attributes on dietary decision-making, while taking into account the effects of attention, and to investigate if there are difference between individuals with and without obesity. We study two aspects of decision-making, the estimation of value by asking participants to rate their willingness to eat an item at the end of the experiment, and the choice of participants by asking them to choose between two items. We also ask participants to rate the nutritional content of these items, to see if their valuation is based on their beliefs about the nutritional content or the true nutritional content.

5.3. Results

We recruited 25 patients referred for obesity treatment and 25 lean controls (Figure 5.1a). To limit the potential confound of nutrition literacy, each participant carried out a nutrition course before taking part in the experiment (see Methods). During the lab visit, the participants carried out three tasks. In a first task (Figure 5.1b), they were asked to rate how much they would like

to eat different snack items at the end of the experiment. We refer to these ratings as subjective value (SV) ratings. In a second task (Figure 5.1c), they were asked to estimate nutritional attributes (fat content, protein content, added sugar content and healthiness) of the snack items. In the last task (Figure 5.1d), the participants were shown on each trial two food items, and had to select the one they wanted to eat. This task was fully incentivized, as participants were informed that they would receive one of the chosen items at the end of the experiment.

The snack items were easy to consume items from the two leading supermarkets in Switzerland and selected to have large variance in their attributes and low correlation between attributes (see Methods and Supplementary Figure 5.1). The pictures of the snack items presented to the participants contained a plate with one portion of the item and the packaging of the item in the background to clarify the nature of the item.

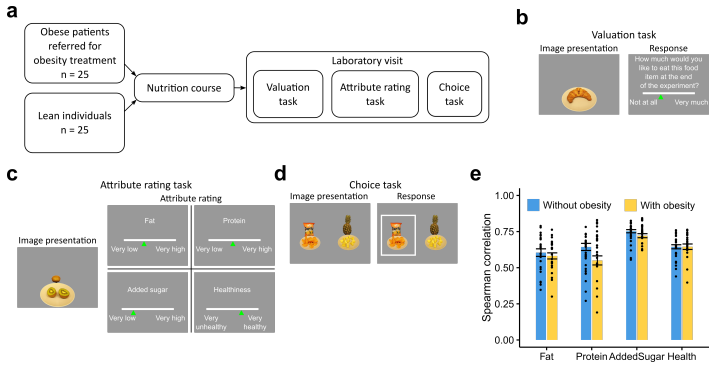


Figure 5.1. Study overview. a) Outline of the study. 25 patients with obesity and 25 lean individuals were recruited. Each participant completed a nutrition course before the lab visit. During the visit, they carried out three tasks. **b)** In the valuation task, they rated how much they were wanted to eat different snacks at the end of the experiment. **c)** In a second task, they estimated the fat content, protein content, added sugar content and healthiness of the snacks. **d)** In the choice task, they choose on each trial the snack they wanted to eat between two alternatives. **e)**. The Spearman correlation between the attribute ratings and the objective values were all positive for the group with (yellow, $t(24) > 13.78, p < 0.001$) and without (blue, $t(24) > 18.15, p < 0.001$) obesity, indicating both groups had knowledge about these attributes. An ANOVA indicated no significant difference between the groups for all attributes ($F(1, 48) = 3.63, p = 0.063$) nor specific attributes ($F(3, 144) = 2.14, p = 0.097$), indicating that there is not significant difference in knowledge about the attributes across groups.

Participants have good knowledge of the nutritional attributes

We first wanted to check if the participants had a good knowledge of the nutritional attributes. To quantify this knowledge, we computed the Spearman correlation between their attribute ratings and the objective per portion attribute values (Figure 5.1e). The Nutri-Score [215] was used as an objective measure of healthiness. These correlation values were positive, indicating that the participants had some knowledge about the attributes of these snacks. We also check if there were difference in knowledge about the attributes between groups. An ANOVA showed no significant difference between the groups for all correlations ($F(1, 48) = 3.63, p = 0.063$) nor between the groups for specific attributes ($F(3, 144) = 2.14, p = 0.097$), which suggests that the differences between groups presented below are not driven by differences in nutrition literacy.

Subjective attribute ratings better explain SV ratings than objective attributes values

Having established that there were no significant difference in the knowledge of nutritional attributes between groups, we next investigated the responses of the participants to the valuation task. In particular, we fit the subjective value ratings of the participants with a hierarchical linear regression (see Methods). We compared the fit of the individual attribute ratings to the fit of objective attribute values per portion or density (as the Nutri-Score is only a density measure and not defined per portion, it was used in both attribute per portion and attribute density models). We compared the goodness of fit of these models with the leave-one-out information criterion (LOOIC) [134]. We found that subjective attribute ratings provided a better fit than objective per portion ($\Delta\text{LOOIC} = 830$) and density ($\Delta\text{LOOIC} = 775$) attributes (Figure 5.2a). Notice that the participants carried out the subjective value rating task before the attribute rating task, therefore they were not influenced to used their attribute ratings for their subjective value estimation by the experiment structure. This indicates that participants base their SV ratings on their beliefs of the nutritional attributes rather than the true nutritional attributes.

SV ratings are best explained by nutritional and non-nutritional attributes

In order to further investigate the role of nutritional and non-nutritional attributes in value estimation, we tried to improve the fit of the model

further by including objective attributes values (either per portion or density) for which no ratings were collected (carbohydrate, sugar, salt, fiber, energy and price), the weight of the portion and low-level image attributes (image colorfulness, entropy, variance and saturation, see Methods). These attributes improved the model fits, with the density attributes providing a better fit than the per portion attributes (ΔLOOIC = 55). Finally, following previous reports [197, 205], we included the interaction of carbohydrate and fat density which improved the model further (ΔLOOIC = 114). This model provided the best fit in both groups separately (Supplementary Figure 5.2). Importantly, the LOOIC metric accounts for model complexity. Therefore, these results cannot be explained by lack of penalization of model complexity and indicate that participants base their SV ratings on many different nutritional and non-nutritional attributes.

No significant difference between groups in the influence of nutritional and non-nutritional attributes on SV ratings

We next investigate how these different nutritional and non-nutritional attributes influence SV ratings for each group. This is captured by the weights of the linear regression (Figure 5.2a). The nutritional attribute with the highest influence was the healthiness ratings ($P_{MCMC} < 0.001$, d = 4.59 and $P_{MCMC} < 0.001$, d = 7.02 in the groups with and without obesity, respectively), and the non-nutritional attribute with the highest influence was image colorfulness (i.e., a metric that correlates with human report of image colorfulness, $P_{MCMC} < 0.001$, d = 5.32 and $P_{MCMC} < 0.001$, d = 4.05 in the groups with and without obesity, respectively). Importantly, we find no significant difference in weights between the two groups, indicating that nutritional and non-nutritional attributes influence SV ratings in a similar way in individuals with and without obesity.

SV ratings but also other attributes influence choice

We continued our investigation by looking at the choices of the participants in the choice task. Our first approach was to fit the choices of the participants with a hierarchical logistic regression (see Methods). We first compared a model considering only the SV ratings to multi-attribute models considering SV ratings and nutritional attributes. If the participants only relied on their SV ratings to guide their choices, we would expect the model considering only SV to have the best LOOIC. However, we found that models considering SV and nutritional attributes outperformed the models considering only SV

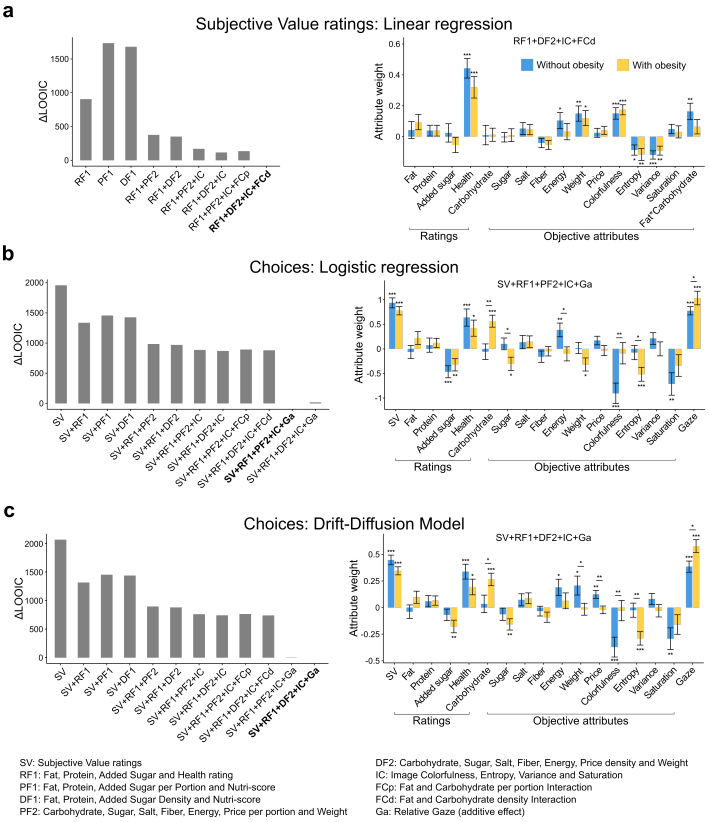


Figure 5.2. Model comparison and model parameters Each panel shows the difference in LOOIC between models (left) and the regression weights of the best fitting model (right). **a)** Model fits of SV ratings with hierarchical linear regressions. Attribute ratings explain SV ratings better than objective attribute values. The best fit is obtained by combining attribute ratings, attributes values per density for which not ratings were collected, portion weight, low-level image attributes and the interaction of fat and carbohydrate per density. In the lean group, we observed a positive influence of healthiness rating, energy density, portion weight, image colorfulness and interaction of fat and carbohydrate density, as well as a negative influence of image entropy and image variance. In the group with obesity, we observed a positive influence of healthiness rating, portion weight, image colorfulness and a negative influence of image entropy and image variance.

Figure 5.2:
(continued) **b)** Model fits of choices with hierarchical multi-attribute logistic regressions. Models combining SV ratings and attribute ratings provided a better fit than models combining SV ratings and objective attribute values. The model fit was obtained by combining SV ratings, attribute ratings, objective attributes per portion for which no ratings were collected, low-level image attributes and relative gaze of the participants. Choices were positively influenced in the lean group by SV ratings, health ratings, energy per portion and relative gaze, as well as a negative effect of added sugar rating, image colorfulness and image saturation. In the group with obesity, we observed a positive influence of SV ratings, healthiness ratings, carbohydrate per portion and relative gaze, as well as a negative influence of added sugar rating, sugar per portion, portion weight and image entropy. We observed differences in weights between groups in the influence of carbohydrate, sugar and energy per portion, image colorfulness and entropy, and relative gaze. **c)** Models fits of choices with a DDM. We observe a similar pattern of model goodness of fit, however the best model considers the attribute density of the attributes for which no ratings were collected. This lead to some changes in the significance compared to the logistic regression. In the lean group, in addition to the effects found in the logistic regression, we observed a positive effect of portion weight and price per density and did not observe a significantly negative effect of added sugar rating. In addition, we did not observe significant differences between the groups in sugar and energy density, but we did observe differences in portion weight and price density. Significant stars indicate significance levels of a Bayesian two-tailed t-test (*$P_{MCMC} < 0.05/2$, **$P_{MCMC} < 0.01/2$, ***$P_{MCMC} < 0.001/2$, non-significance is not indicated).

ratings (ΔLOOIC > 502, Figure 5.2b), which indicates that the participants do not only rely on their SV ratings to guide their choice, but are also influenced by other attributes.

Choices are better explained by subjective attribute ratings than objective attributes values

We next investigated if participants based their choices on their belief about the nutritional attributes or the true nutritional attributes. Mirroring the results of the SV ratings models, we found that a model considering attributes ratings outperformed the models considering true attribute values (ΔLOOIC > 90). In addition, the model could be further improved by adding the attributes for which no ratings were collected and low-level image attributes (ΔLOOIC = 450 and ΔLOOIC = 467 for per portion and density attributes, respectively). Including the interaction of fat and carbohydrate did not improve the model (ΔLOOIC = -9). To investigate the effect of attention on choice, we added the relative gaze (i.e., how much time the participant looked at the left item relative to the right item (see Methods)) as a regressor, which led to the best fitting model (ΔLOOIC = 869). These results indicate that, as SV ratings, the choices of participants are based on the participants

beliefs about the attributes rather than the true attributes, that these choices are influences by many nutritional and non-nutritional attributes and the relative gaze of the participants.

Nutritonal and non-nutritional attributes influence choice differently in individuals with and without obesity

We then investigated the weights of the logistic regression. We report these weights in (Figure 5.2b). The SV ratings had the highest effect on choice ($P_{MCMC} < 0.001$, d = 9.24 and $P_{MCMC} < 0.001$, d = 9.23 for individuals with and without obesity, respectively), indicating that choices are strongly driven by SV estimates. However, as described above, the nutritional and non-nutritional attributes also have an influence, showing that participants incorporate additional information that was not considered during the SV ratings task.

Contrasting with the results of the SV ratings, we do find differences in attribute weights between the two groups, which indicates that individuals with obesity are influenced differently by these attributes. The nutritional attribute with the largest difference between groups was carbohydrate per portion, which had a higher influence on choice in the group with obesity than the group without ($P_{MCMC} = 0.001$, d = -3.10). We also found significant differences in the influence of sugar per portion ($P_{MCMC} = 0.012$, d = 2.25) and energy per portion ($P_{MCMC} = 0.007$, d = 2.44). We also observed differences in the influence of non-nutritional attributes, in particular image colorfulness ($P_{MCMC} = 0.003$, d = -2.69) and image entropy (i.e., how much variation there is in pixel values, $P_{MCMC} = 0.019$, d = 2.18).

Accounting for reaction times leads does not change the main results of the logistic regression

It has been suggested that response times (RTs) carry meaningful information about the decision process. A popular modal to account for RTs is the drift-diffusion model (DDM) [143, 216]. The DDM assumes that evidence for one item versus the other is accumulated over time until a bound is reached which causes the decision maker to choose the corresponding option. The rate of the evidence accumulation depends on the product of the difference in attribute values and the corresponding weights that are fit, in contrast to the logistic regression where the difference in attributes and weights directly influence the choice probability. This change allows the DDM to not only account for the choices but also the RTs of the participants.

We investigated if accounting for RTs would change our results. We repeated our analysis of the choices of participants using a DDM. The best fitting model considered the same attributes than the logistic regression, however the attributes for which no ratings were collected were considered by density instead of per portion (Figure 5.2c). The weights of the DDM were mostly similar to the ones of the logistic regression (Figure 5.2c) and the differences can likely be explained by switching between using attributes values for which no ratings were collected per portion or density. Overall, besides the change from per portion to density values, accounting for RTs does not change the main results of the logistic regression. Models considering attribute ratings instead of attribute true values provide a better fit, participants base their choice on many nutritional and non-nutritional attributes, and these attributes affect differently individuals with and without obesity.

Eye-tracking reveals difference influence of overt attention on choice

In both the logistic model and DDM, we considered the influence of gaze. There are two approaches to include gaze. Gaze can be considered to have an additive effect (i.e., there is a bias towards the item that is looked at longer, in which case a regressor is simply added to the model) or a multiplicative effect (i.e., gaze influences how the other attributes influence choice, in which case a discount factor is used to model how much the attributes of the fixated item influence choice compared to the attributes of the unfixated item). Previous reports have shown that a multiplicative model better captures human behavior than an additive model [217]. We tried both approaches. In the logistic regression, only the additive model reached convergence. In the DDM, the additive model outperformed the multiplicative model (ΔLOOIC = 93). Therefore, our results are not consistent with the literature. However, notice that typically these models do not contain as many regressors, which could explain this discrepancy. Nevertheless, we investigated the effect of gaze based on the additive approach. We found that the relative gaze had a positive influence on choice ($P_{MCMC} < 0.001$, d = 9.14 and $P_{MCMC} < 0.001$, d = 7.45 for individuals with and without obesity, respectively) in the logistic regression, consistent with previous reports of a correlation between looking at an item and choosing the item [64, 65]. In addition, the gaze affected differently individuals with and without obesity ($P_{MCMC} = 0.009$, d = -2.40), with individuals with obesity having a higher correlation between looking at the items and choosing them. This results indicates that not only do the influence of nutritional and non-nutritional attributes on

choice differs in individuals with obesity, but also the effect of overt attention on choice.

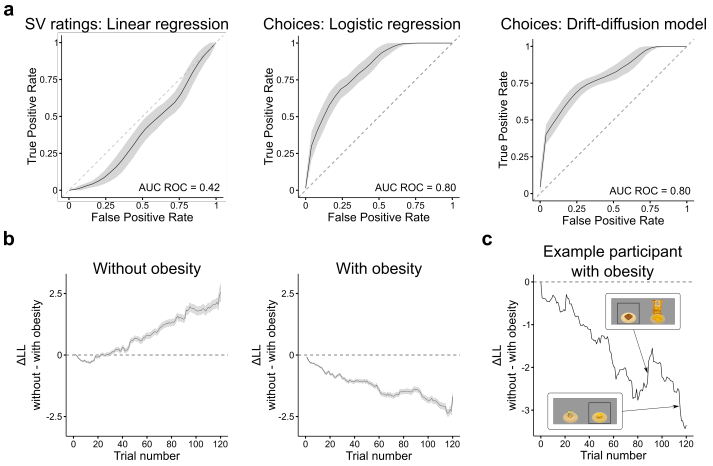


Figure 5.3. Model predictions. a) Receiver operating characteristic (ROC) curve of the linear model explaining SV ratings (left), logistic model explaining choices (middle) and DDM model explaining choices (right). Shaded areas represent standard deviation computed by bootstrapping across leave-two-out permutations. The grey dashed line represent a random model. The model explaining SV ratings does not predict accurately the BMI group of the participants. However, the models explaining choices do perform above chance level. **b)** Average difference in log-likelihood (ΔLL) over trial number for participant without (left) and with obesity (right). Positive values represent a model belief that the participant is lean, and negative values represent a belief that the participant has obesity. The difference in ΔLL between the groups increases with trial number, indicating that the model increases its prediction confidence. **c)** Evolution of ΔLL for an example participant with obesity. On trial number 89, the participant chose a bread roll over tortilla chips, which increased the model's belief that this participant does not have obesity. On trial 115, the participant chose vanilla custard instead of fruit yogurt, increasing the model's belief that this participant has obesity. AUC-ROC: area under the receiver operating characteristic curve.

Choices but not SV ratings predict the BMI group of participants

As we have used simple multi-attribute models, they can easily be interpreted (e.g., identifying that the influence of colorfulness on choice is different between individuals with and without obesity). Therefore, if the models could predict the BMI group of the participants, we could not only use them to classify an individual, but also to identify which factors of this individual are more similar to individuals with or without obesity. With this in mind, we investigated if our models could accurately classify the BMI group of participants based on

their responses. To this end, we conducted a leave-two-out cross validation approach, where we fit the model on the data excluding one participant from each group. Using our hierarchical framework, we derived the population weights for each group. We then used these estimates to compute for each trial the log-likelihood that the left-out participants have obesity ($LL_o(p,t)$, where p corresponds to the participant number and t is the trial number) and the LL that the participants are lean ($LL_l(p,t)$), given their responses. We can then compute the belief that the participant p is lean given their responses on trial t:

$$\delta LL(p,t) = LL_l(p,t) - LL_o(p,t) \tag{5.1}$$

If $\delta LL(p,t)$ is positive, this indicates that the model believes the response on trial t of participant p to be more consistent with a lean than an obese BMI, and the opposite if the value is negative. The higher the value, the more confident is the model. To decide how the participant should be classified after completing T trials, we can sum:

$$\Delta LL(p,T) = \sum_{t=1}^{T} \delta LL(p,t) \tag{5.2}$$

If we set T to the total number of trials T_{max}, then $\Delta LL(p,T_{max})$ captures our belief about the BMI group of the participant after completing all trials. If this value is positive, the model believes that this participant has a lean BMI and the opposite if the value is negative. We may choose to bias our model towards classifying participants as belonging to a lean or obese BMI group, for example by requiring the threshold to be classified as lean to be 2 instead of 0. Changing this bias influences the sensitivity (true positive rate) and the specificity (true negative rate). Depending on the goal, it may be useful to bias the model one direction or the other. However, in our case we are simply interested in the predictive performance of the model. To measure the predictive performance of the model, we measure the AUC-ROC metric, which captures performance for different levels of trade-off between sensitivity and specificity. The AUC-ROC metric takes a value of 1 for a perfect classification and 0.5 for a chance-level classification.

The model explaining the SV ratings of the participants was unable to distinguish between groups (AUC-ROC = 0.42, Figure 5.3a). This was unsurprising, as we did not find any difference in the weights of the model explaining SV

ratings. However, the models explaining the choices correctly classified individual with and without obesity (AUC-ROC = 0.80 for both the logistic and the DDM model). Figure 5.3b illustrates how the predictions of the logistic model evolve over trial number for each group. As the number of trials increases, the difference in classification likelihood between the groups increases, indicating the that model increases its confidence. After around 50 trials, the two curves are clearly distinguishable, suggesting that the model has already correctly classified individuals. An advantage of our approach is the ability to identify which choices changed the beliefs of the classifier. As an example, we show in Figure 5.3c the time course of the predictions of the logistic model for a single participant and highlights which choices changed the belief of the model. For example, on trial 115, the participant chose vanilla custard over a fruit yogurt, increasing the model's belief that this participant has obesity. Taken together, these results clearly indicate that choices carry information about the BMI group of the participants, but the SV ratings do not.

5.4. Discussion

Our main result is that choices, but not SV ratings, are different between participants with and without obesity. These choices are influenced by SV ratings, nutritional and non-nutritional attributes, and overt attention (as captured by the relative gaze). We built a classifier based on our multi-attribute model, which can not only predict the BMI group of participants, but is also easily interpretable, in terms of which choices and which attributes weight on the prediction. We achieved an AUC-ROC of 0.80, which is surprising given the simplicity of ours models and paradigm. This accuracy may be improved by adding more participants, more food items or more attributes (see discussion below). One may also use more sophisticated machine learning methods such as neural networks, however their downside is that they are not easily interpretable, therefore understanding which attributes are responsible for the differences between groups is difficult to understand. Additional work is needed, but overall, a classifier could be used in the future by clinicians to follow the behavior phenotype of their patients throughout a treatment or to identify individuals susceptible to develop obesity. For this application, building interpretable models as we have done will be valuable.

We observed that both groups do not only base their choices on their SV ratings, but are in addition influenced by other nutritional and non-nutritional attributes. This result is reminiscent of preference reversals observed in risky

choice [218]. Participants are willing to bid more for risky gambles, but choose low-risk gambles in a choice task. This result can be explained by changes in attention to different attributes between auction and choice tasks as revealed by eye-tracking [219, 220]. One may speculate that a similar process is at play to explain difference between SV rating and choice in our study. Although eye-tracking cannot be used to disentangle the attention to different attributes in our food choice task, our computation models can. For instance, they indicate a positive influence of image colorfulness during SV rating but a negative influence of image colorfulness during choice in the lean group. This suggests that the context of the task changed the attention of participants to colorfulness. We cannot rule out the possibility that the participants changed their preference between the rating phase and the choice phase, which could be expected as participants become more hungry. Future studies could collect SV ratings after choices to control for this. However, it has been suggested that participant change their ratings based on their choice [221], which makes disentangling the effect of hunger a challenging question.

A result of this work is that participants base their dietary valuation on their beliefs about the nutritional attributes for fat, protein, added sugar content and healthiness rather than the true values of these attributes. This is consistent with a previous report showing that a model considering ratings for fat, carbohydrate, sugar, protein, sodium and vitamins [196] performed better than a model considering the true values of these attributes. However, the opposite was shown in the case of calories [195]. It is unclear if calories weight differently on valuation than other attributes, or if this difference is simply due to difference methodologies or difference in the study group. Further investigating the difference between calories and other nutritional attributes could be a line of work for future research.

We found an influence of low-level image attributes in dietary decision-making. These could be due to a link between flavor and low-level image attributes in the food items (e.g., chocolate flavored items are less colorful than fruit flavored items). Another view is that low-level image attributes directly influence the participants' valuation. For instance, it is known that color can influence the perception and preference of foods [222]. Future studies could manipulate these low-level image attributes to understand their effect on valuation. We also observed difference between the groups in the influence of these low-level image attributes in choices. Intriguingly, obesity has been associated with deficits in visual cortical function [223]. Future studies could

investigate if these deficits are linked to differences in the influence of low-level image attributes during choice.

Consistent with previous reports [64, 65], participants chose more often the items that they looked at longer. This effect was larger in participants with obesity, as if they were ignoring more the option that they are not looking at. This could be linked to the cognitive deficits observed in obesity, in particular in working memory [224, 225]. Future eye-tracking studies could investigate this question. The underlying causality of this effect is also unclear. Although it has been shown by manipulating attention that participants tend to choose items that they look at longer [226], therefore suggesting that this is effect is not driven by participants looking longer at the item they want to choose, it is unknown if this effect is also driving the difference between participants with and without obesity.

We decided to model subjective value ratings and choice with a linear and logistic regression respectively. In doing so, we implicitly assumed that the value is constructed from a linear combination of nutritional (and non-nutrientional) attributes. With the exception of fat and carbohydrate, we did not investigate the interaction of attributes. However, the enthusiasm for cooking and eating complex foods in modern societies suggests that food is more than the sum of its attributes. In addition, seminal food science work has shown that attributes have a quadratic effect on pleasure [227, 228]. For instance, sweetness increases liking until a certain point after which further increase causes a decrease in liking. In addition, food choices can also be influenced by the decision-maker's beliefs about the social and environmental impacts of the foods [229] as well as social norms [230, 231]. This complexity is ignored by our current model and could be the subject of future work. Different solutions have been proposed to deal with this complexity. One approach is to use large-scale crowd-sourced data to identify which attributes are meaningful for choice [232]. Another approach suggests that the brain does not combine attributes to construct value but rather compares options to previously memorized experiences [233]. Overall, a better understanding of how the brain processes value will likely improve our understanding of obesity [234, 235].

Many studies investigate dietary decision-making as a way to understand value based decision-making [74, 122, 236–238]. Similar to our study, a common paradigm is to ask participants to rate their willingness to eat different items, and then to ask them to choose between two items which one they want to consume. It has previously been shown that choices can be better explained

by adding attribute ratings in addition to the subjective value ratings, such as taste, healthiness, texture and appearance [203]. Our results echo these findings and indicate that simply adding objective attribute can significantly improve the fit of choice data. In addition, our results indicate that the Nutri-Score [215] can be used as a proxy for healthiness ratings. Together, these results suggests that future decision-making studies could add objective attributes to improve their model fits and thus their ability to test their hypotheses.

Our study does not provide any knowledge about the causality of the effects. For instance, does obesity cause individuals to choose more the item that they look at longer, or is this caused by a factor of obesity? Although age and gender were similar across groups, we cannot exclude that our results are linked to another factor which is correlated to obesity. Longitudinal studies are necessary to answer these questions. Our study is also limited by the fact that is does not tackle the questions of hunger and blood nutrient levels. Although participants were instructed to not eat before the study, it is unknown if both groups reacted similarly to fasting and had the same level of hunger and blood nutrients. Finally, a common limitation of dietary decision making studies is that it is unclear if these results are dependent of the food item set or cultural preferences. Reproducing these results in another country would improve the their generalizability.

It has been suggested that psychiatric disorders can be better understood by phenotyping human behavior through computational modeling [239]. We believe a similar approach is possible for obesity [240] and that the presented work for is a step in this direction. Future studies could combine this behavior phenotyping with neural imaging data [241] to further identify the latent parameters associated with obesity. Understanding these latent parameters could help tailor the prevention and treatment of obesity.

5.5. Methods

Participants

We recruited 25 patients (17 females, 42.9 years old ± 12.1, BMI 38.3 ± 4.7 (mean±SD)) who were referred for a clinical obesity intervention and 25 lean controls (16 females, 41.4 years old ± 13.4, BMI 21.9 ± 2.4 (mean±SD)). We excluded participants who did not fit MRI inclusion criteria (e.g., claustrophobia). One participant could not carry out the task as they did not fit comfortably in the MRI scanner, so another participant was

recruited to replace them. All participants provided informed consent before participating in the study and were compensated with 100CHF for their participation. The study conformed with the declaration of Helsinki and was approved by the ethics committee of the Bern canton.

Stimuli

The stimulus set consisted of 64 pictures of food items. The pictures contained one portion of the item on a plate and the package of the item in the background. The portion size was based on the package information when available. To select the food item set, we identified 272 food items from the two leading supermarkets in Switzerland that we considered could be eaten as a snack. We selected a subset of 64 food items that minimized correlations between fat, protein and sugar content and the Nutri-Score. For each participant, we replaced the items that they did not know or could not eat due to dietary restrictions which a similar item for which we had images. This lead to the replacement of an average of 3.2 images for 6 patients and 1.2 images for 5 controls. Nutritional attributes were collected using the package information when available and from Open Food Facts (ch.openfoodfacts.org) when necessary. The Nutri-Score was computed using an online calculator (simonetthomas.github.io/CalculateurNutriscore/index.html) using the general category for all items. We used the numerical values from -15 to 40 instead of the 5 letter categories for a more refined description of healthiness. We multiplied these values by -1 for the results presented in Figure 5.1 and Supplementary Figure 5.1, such that positive values reflected higher healthiness to be consistent with the participants' ratings. We extracted the image entropy and variance in R with the glcm package [242] and the image saturation with the magick package [243]. Image colorfulness was computed with the efficient colorfulness computation proposed by Hasler and Süsstrunk [244]. We chose to limit our analysis to these low-level image attributes as they had a relatively low correlation between each other (Supplementary Figure 5.1).

Behavioral paradigm

To reduce the potential inter-individual differences in nutrition literacy, all participants completed an online nutrition course which lasted about 25min before visiting the lab. To minimize the difference in hunger level we asked participants not to eat 4 hours before coming to the experiment. Prior to the tasks, the participants were shown all images of food items and asked to report items that they did not know or could not eat due to dietary

restrictions. They were informed that all question referred to the portion size that was presented on the plate in the image. The behavior paradigms were run using Pschychtoolbox [245–247] on Matlab 2019a (Mathworks, Natick, MA, USA).

Valuation task

The valuation task was performed in the MRI scanner. On each trial, the participants were shown an image for 3s. They were then asked to answer the question "Wie gerne würden Sie dieses Lebensmittel am Ende des Experiments essen?" ("How much would you like to eat this food at the end of the experiment?") by moving a cursor on a response bar going from "gar nicht" ("not at all") to "sehr gerne" ("very much"). They could move the curse left or right using two different buttons and confirm their response with a third button. The initial position of the cursor was randomized on each trial. The participants had no explicit time limit to respond but were asked to answer faster than 5s on average. Their response was then shown on the screen for 0.5s. The intertrial interval was drawn from a gamma distribution with shape parameter 6 and scale parameter 1 and truncated between 2 and 15s. The participants completed 8 runs of 16 trials each, thus each item was rated twice.

Attribute-rating task

The attribute-rating task was performed outside the MRI scanner. The participants were asked to rate the items based on their fat content ("Wie hoch ist der Fettgehalt dieses Lebensmittels?"), their protein content ("Wie hoch ist der Proteingehalt (Eiweissgehalt) dieses Lebensmittels?"), their added sugar content ("Wie hoch ist der zugesetzte Zuckergehalt dieses Lebensmittels?") and their healthiness ("Wie gesund schätzen Sie dieses Lebensmittel ein?"). On each trial, the item was presented for 3s. The participants were then asked to rate the items by moving a cursor on a response bar from going from "sehr niedrig" ("very low") to "sehr hoch" ("very high") or "sehr ungesund" ("very unhealthy") to "sehr gesund" ("very healthy") in the case of health ratings. The initial position of the cursor was randomized on each trial. The participants could use a mouse to move the cursor and a left click to respond. The participant had no time limit to respond. Their response was shown on the screen for 0.5s, followed by an intertrial interval jittered between 0.5 and 1s. The participants rated the different attributes one after the other in a random order. For each attribute rating, the images were presented

in a random order. In both the attribute-rating task and the choice task, the participants sat about 70cm from a 24inch monitor (Philips brilliance 240b).

Choice task

The choice task was performed outside the MRI scanner. On each trial, two food items were displayed simultaneously, one on the left and one on the right side of the screen. The participants were asked to choose the item they wanted to eat at the end of the experiment. To incentivize them to respond truly, they were informed that one trial would be randomly selected at the end of the experiment and that they would receive the item they chose on that trial. There were given a time limit of 5s to respond and could respond using the left and right arrow keys of the keyboard. In total the participants missed the response deadline 69 times. Their response was kept on the screen for 1s. The task had an intertrial interval jittered between 0.75 and 1.25s.

To select the items pairs presented in the choices we computed the average SV rating of each item based on the ratings of the valuation task and the SV difference of each item pair. We divided the SV difference in 4 levels (0%-5%, 5%-10%, 10%-15% and 15%-20% of the length of the rating scale). The trials were fully counterbalanced across these differences in subjective value levels as well as the location of the higher SV option (left or right).

Eye-tracking

Eye-tracking data was collected for 20 participants in the group with obesity (14 females, 49.7 years old $\pm$ 13.2, BMI 37.6 $\pm$ 4.6 (mean$\pm$SD)) and the group without obesity (12 females, 36.0 years old $\pm$ 9.5, BMI 21.9 $\pm$ 2.6 (mean$\pm$SD)) using the Tobii Pro Nano (Tobii Pro AB, Stockholm, Sweden) and the Titta toolbox [248]. The eye-tracker was calibrated for each participant using 5 calibration points. We considered that the participant was looking at a given item if his gaze was more than 5cm towards that item compared to the center of the screen. Trials in which the gaze contained missing values were removed (31 trials from participants without obesity and 2 trials from participants with obesity).

Behavioral analysis and statistics

Model description

The weights (including the intercept) of the linear regression, logistic regression and DDM were all fit hierarchically. Each participant level parameter was drawn from a group distribution. The group distribution mean was specified with a normal prior with mean 0 and precision 0.1 and standard deviation was specified with a uniform distribution from 0.0001 to 100. The precision of the normal distribution was chosen to regularize the parameter estimates, thus addressing issues of multiple comparisons [249, 250]. In the linear regression, we specified for each group the standard deviation of the error between the model fits and the data with a uniform distribution from 0.0001 to 100. This parameter was share by both groups in the leave-two-out analysis. In the DDM, the bias, non-decision time and boundary separation were fit hierarchically. The group level bias means were specified with uniform priors from 0.01 to 0.99 and standard deviations specified with uniform priors from 0.01 to 1, the group level non-decision time means were specified with uniform priors from 0.01 to 1 and standard deviations specified with uniform priors from 0.01 to 1, and the group level boundary separation time means were specified with uniform priors from 0.01 to 5 and standard deviations specified with a uniform prior from 0.01 to 10. To improve the stability of the DDM, the drift rates were drawn from a distribution with a standard deviation specified for each group with a uniform prior between 0.1 and 5. This parameter was shared by both groups in the leave-two-out analysis. The drift rates means were computed by multiplying the attribute weights with the attributes values (without an intercept). All attributes were z-scored for each participant individually, therefore we assumed that the participants did not share the same representation of the rating scale. In all models containing information about the gaze, the gaze attributes only affected the parameters from the participants that had eye-tracking data collected.

Statistical inference

Posterior inference of the parameters in the hierarchical models was performed via the Gibbs sampler using the Markov Chain Monte Carlo (MCMC) technique implemented in JAGS [251] using the runjags [252] and rjags packages [253]. The chains were run with a minimum burn-in parameter of 4 000 (the first 4 000 samples were ignored to ensure that the starting sample did not influence the results). A minimum of 10 000 samples were then drawn, however only 1 in 5 samples were kept for the final inferences, reducing

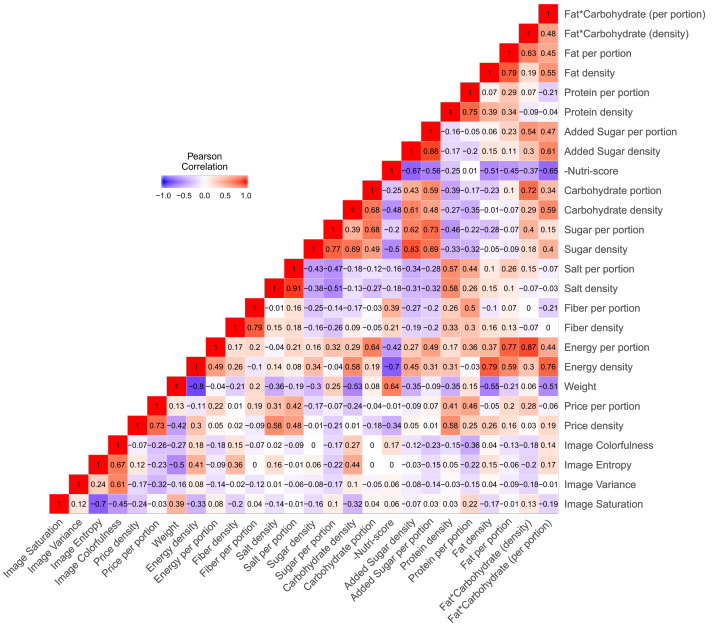
the auto-correlation of the samples. This computation was done for three independent chains (using difference random number generators and different starting samples). We conducted Gelman–Rubin tests for each parameter to confirm the convergence of the chains. All latent variables in our Bayesian models had $\hat{R}$<1.05, which suggests that all three chains converged to a target posterior distribution. We checked via visual inspection that the posterior population-level distributions of the final MCMC chains converged to our assumed parametrizations. For all random effects reported here, the reported value corresponds to the mean of the standardized posterior distribution, and the "P-values" reported for these regressions are not frequentist P-values but instead directly quantify the probability of the reported effect differing from zero (P_{MCMC}). They were computed using the posterior population distributions estimated for each parameter and represent the portion of the density functions that lies above/below 0 (depending on the direction of the effect).

Data and code Availability

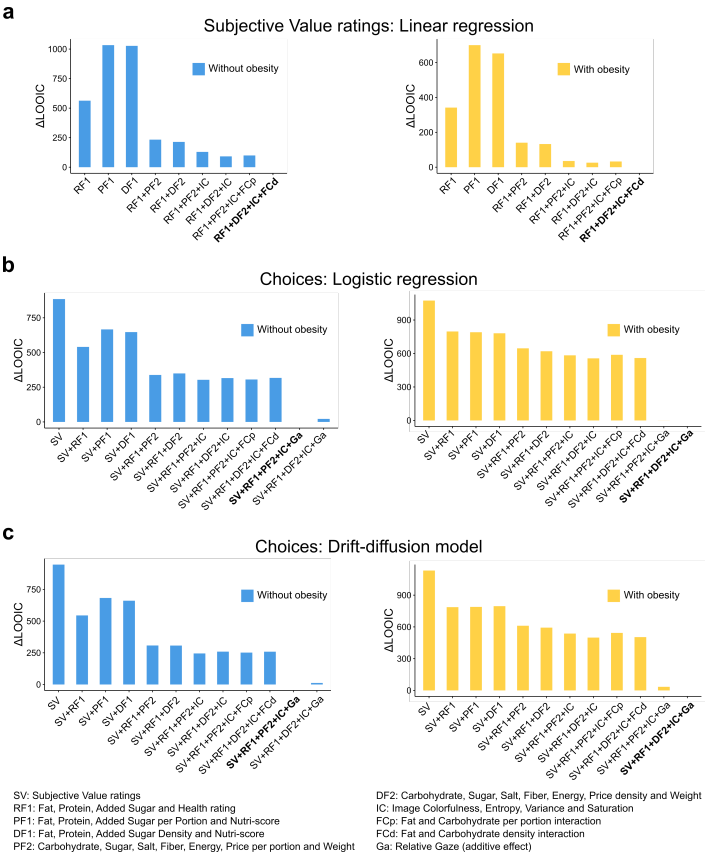
Data and code will be made openly available upon manuscript acceptance.

Acknowledgements

This work was supported by grants to Rafael Polanía and Lia Bally from the Vontobel Foundation and the Johanna Dürmüller-Bol Foundation.



Supplementary Figure 5.1. Correlation of attributes. Pearson correlation of all objective attributes considered in this study. For consistency with the main text, The Nutri-Score values are multiplied by -1, such that higher scores indicate more healthiness.



Supplementary Figure 5.2: Model comparison for each group. **a)** Model fits of SV ratings with linear regressions for the group without (left) and with obesity (right). The best model is the same in both groups, as presented in the main text. **b)** Model fits of choices with logistic regressions for the group without (left) and with obesity (right). The best models have the same level of complexity in both groups, but the attributes per portion perform better in the group without obesity (ΔLOOIC = 22) and the attribute density perform better in the group with obesity (ΔLOOIC = 0.75). **c)** Model fits of choices with a DDM for the group without (left) and with obesity (right). Similar to the logistic regression, the best models have the same level of complexity in both groups, but the attributes per portion perform better in the group without obesity (ΔLOOIC = 12) and the attribute density perform better in the group with obesity (ΔLOOIC = 34).

6

GENERAL DISCUSSION

6.1. Chapter discussion

6.1.1 Chapter 2

In chapter 2, we followed the framework proposed by Marr [126] to study the brain. According to this framework, the brain should be studied at different levels: the computational level (i.e., the goals of the organisms), which correspond in our case to the three different goals of the decision-maker we tested, the algorithmic level (i.e., the algorithms used by the organisms), which correspond in our case to efficient coding, and the implementation level (i.e., how is the algorithm implemented in the brain), which correspond in our case to the perceptron. Our study highlights the importance of considering all these levels and how the principle of efficiency can help constrain the investigator search space. This led to a surprising result, the neural code for numerosity does not simply follow a maximizing accuracy or fitness goal, but rather seems to have evolved to maximize accuracy while economizing on the number of samples used to represent the prior. This goal has the benefit of increasing the ability of the decision maker to adapt to new environments.

The idea of perceptual adaptation is not new. For example, we adapt our luminosity perception to darker environments [254]. This process, known as dark adaptation, is carried out in part by synaptic adaptation of retinal neurons. Our results support previous evidence that a similar adaptation also takes place for higher-level percepts [255, 256] such as numerosity [257] and that this adaptation is taken into account at the computational level. Adaptability is relevant for economic decisions. For instance, it has been shown that people adapt their willingness to pay for food items [85] and their risk seeking behavior [159] after being presented with adaptation trials. These results have both been linked to principles of efficiency and interrogate on the relationship between numerosity and value perception.

Value (or expected utility) is a magnitude, as is numerosity. Value and numerosities may share similar environmental priors. For example, the prices

in a supermarket follow a power-law distribution [81], with cheaper prices occurring more frequently. As previously mentioned, there is an ecological intuition of why the ANS should follow Weber's law, fighting against two instead of one individual makes more of a difference than fighting 31 instead of 30, even if the difference in number of individuals in only one in both cases. However, it is unclear why value perception should follow Weber's law. You may be willing to use a coupon to save a franc when buying cheese but might consider it too much effort when doing a large purchase such as a bike. This behavior is irrational, as in each case the franc saved is the same [258]. This bias could be due to implication of numerosity perception in such value decision-making tasks. Khaw et al. [85] explored the relation between numerosity perception and value further. They explained the variability and risk aversion when choosing between gambles by imprecisions in numerical perception. Neural evidence supporting this hypothesis was found. Individuals with less precise neural magnitude representation as measured by fMRI behaved with more variability and risk aversion when choosing between gambles [259].

An open question is the neural basis of numerosity perception. In our work, we assumed that numerosity was encoded by the average activity of a group of neurons, instead of a population code in which each neuron is tuned to a preferred numerosity. However, there has been some evidence for neurons that respond to a preferred numerosity [83, 84, 260]. It is unclear if these two systems coexist or if we captured in our work principles that were also present in the preferred numerosity code. Importantly, work identifying value representation in the brain is consistent with a neural code based on the average population activity. For instance, fMRI studies typically find that the activity of the vmPFC correlates with the value of the options in a value-based task [261]. Therefore, these two coding schemes may coexist, but apply to different modalities.

6.1.2 Chapter 3

In chapter 3, we found that a sequential encoding and Bayesian decoding model better explains numerosity perception than a thermodynamically inspired model. Notice that both models consider resource limitations, however they do so in a different way. In the SEB model, variability is the results of noise at the encoding level (similar to [262]). The decision-maker decodes this noisy representation optimally (without noise). Whereas in the TIM, the decision-maker pays a cost to move from a default state to a state that carries information about the environment. The decision-maker then samples from

the distribution represented by this new state, so variability is considered at the action level. Therefore, variability is considered at opposing ends of the perception process by these models. One could argue that, as noise if found throughout the brain [8], it should be considered at both stages of the model. The presented models could easily incorporate an additional noise term, however this would reduce their parsimonicity. Future research could investigate a parsimouous model that considers noise throughout the perception process, which may lead to a better understanding of human behavior.

Both models predict that given unlimited resources, the decision-maker should be perfectly precise. However, they make different predictions about the opposite extreme. If the exposure time was a Dirac delta (i.e., infinitely small), the SEB model would simply perceive the average of the prior distribution whereas the TIM model would sample from the prior distribution. Therefore, these models have opposite views on how people behave when they do not have any information: is their educated guess consistent or variable? These different views have important implications for human behavior, especially given the previously discussed relationship between value and numerosity. Consider the following scenario. Every day, you walk in front of an auction house, and you are given the chance to place an bid on a piece of art. You are not an art expert, so you cannot estimate the true value of these pieces of art. How do you behave? Do you consider that each piece of art has an average value and place a bid only if it's below a certain value? Or do you sample each day from your prior for value of pieces of art and decide to place a bid if your sample is above the proposed price. This second approach would predict that if the auction house shows you the same piece of art enough days in a row, you will eventually place any arbitrarily high bid within the prior range, whereas the first approach would predict that this marketing strategy would fail. In other words, the TIM model allows for more extreme behaviors. This discussion underlines an important aspect, there are multiple ways to consider resource limitations [263]. Even if the differences between models seem relatively small, they can be important to study. One major challenge to implement resource rational models is to correctly identify the resource limitations. One of the proposed approaches is to ground these limitations in information theory [264], as we have done in our work.

Notice that no matter the stimulus duration, the coefficients of variation are mostly constant for all numerosities in the behavior data (Figure 3.5). Based on this observation, we would expect numerosity discriminations to follow Weber's law, even if stimuli are presented for a short duration. This contrasts

with work that has found that rats violate Weber's law for sound perception if the duration of the stimuli is limited [47] (see Appendix A). This second view would also predict that discrimination involving larger numerosities would be faster than discrimination involving smaller numerosities. However, the opposite is typically found [265, 266]. As our model incorporates time, it could be well suited to be adapted for a discrimiation task. It could further be adapted to incorporate a continuous magnitude, such as sound intensity, and could be used to compare how time affects numerosity and sound discrimination differently. In chapter 2, we suggested that numerosity perception can be modelled by a different coding architecture than low-level sensory systems (i.e., an average population activity code instead of an individual neural tuning code), while keeping similar principles. Future work could investigate if this difference explains why different sensory systems behave differently when stimulus duration is limited.

6.1.3 Chapter 4

In chapter 4, we uncovered a mechanism of top-down control of non-spatial attention. In particular, we found that the top-down control of non-spatial attention relied on the fluctuation in excitability states of the IFJ which was manipulated to be more or less synchronous to the activity in sensory areas. Importantly, these fluctuations are not an endogenous neural oscillation, but were externally manipulated by our task and stimulation protocol. This result indicates that the communication through coherence hypothesis does not only apply to endogenous oscillations but is a more general principle. Therefore, our protocol provides a new tool to non-invasively stimulate the brain based on the communication through coherence hypothesis. Future work could test the relevance of different frequency bands by tuning the frequency of the task and stimulation. If stimulation at specific frequencies band have higher than expected performance, this would indicate that the brain is tuned to these specific frequency bands, which would likely correspond to endogenous oscillations. Given that neural oscillations have been implicated in many cognitive tasks [102, 267, 268], future work could apply our stimulation protocol to test the causality of fluctuations of excitability states, and eventually the causality of endogenous neural oscillations. Importantly, differences in neural oscillations have also been observe in cognitive disorders [269–271]. Therefore, non-invasive brain stimulation (based on our stimulation protocol or a traditional protocol) could be used as a therapeutic tool. It has already been shown that tACS can increase working memory capabilities in older adults [269–271]. We have discussed that neural oscillations are compatible

with an efficient coding scheme. However, these results have only been shown for oscillations in the gamma frequency band. Neural oscillations are typically found in different frequency bands and these oscillations have been shown to interact across frequencies, an observation termed cross-frequency coupling. It has been suggested that this cross-frequency coupling allows the brain to coordinate its activity to communicate efficiently [272]. However, this work has not been in connected to efficient coding mechanisms. Future modeling work could aim to bridge this gap.

6.1.4 Chapter 5

In chapter 5, we applied our framework to study a disorder related to decision-making, obesity. We found differences between individuals with and without obesity in the influence on choice of different attributes and of gaze. The influence of these attributes on choice may be similar to the process of non-spatial attention discussed in chapter 4. Future research could aim to stimulate individuals with obesity in a similar way than the methods employed in chapter 4 to enhance or disrupt the influence of different attributes in the valuation process to make them similar to the influence in the healthy group. If the influence of different attributes in the valuation process is an underlying cause of obesity, this approach could serve as a treatment. We observed that individuals with obesity were more likely to choose the item that they looked at longer. The causality of this effect remains unclear, but it could be tested in future experiments, for example by varying the presentation time of different items. Importantly, if individuals with obesity are more likely to choose items that they look at longer, they may be more susceptible to marketing strategies that show items to potential customers. If this hypothesis is true, policy makers should consider regulating the marketing strategies. For example, the Chilean government passed a food labeling law requiring among other provisions that items high in calories, sugar, sodium and fat be marked with a warning label and banning advertisement of these products to children. This has led to a decrease in high-sugar beverage consumption [273], suggesting that regulation may be a valid approach to reduce the obesogenicity of our environment.

6.2. General discussion

In this thesis, we separated the *when*, the *what* and the *how* of neural efficiency. However, this separation is arbitrary and multiple aspects may be considered simultaneously. For instance, pupil size correlates with variability

in perceptual decision-making [274]. It has also been found that numerosity estimation biases increase when participants are required to do a distractor task [138]. These results indicate that arousal and attention can change the amount of resources used in the encoding-decoding process. In other words, the *when* and the *what* of neural efficiency can interact with the *how* of neural efficiency. A different line of work is investigating how arousal and attention are related [275]. Together, this suggests that the multiple aspects of efficiency can be considered together to build a more complete understanding of the brain. Throughout this thesis, we have discussed how the principle of efficiency can help build models of the brain and decision-making. We have applied these principles to study numerosity perception, non-spatial attention and obesity. This thesis has demonstrated that incorporating the principle of efficiency in our models is a promising approach to better understand the brain and decision-making and could help better understand decision-making disorders, inspire developments in artificial intelligence and provide a better understanding of what we are.

Considering neural efficiency principles has already inspired advances in artificial intelligence. For example, attention mechanisms have been included in neural networks [276]. Instead of considering all input data equally, a network with attention will focus more on specific parts of the data. This is particularly relevant for tasks in which the input and output have sequential structures, such as translation, object recognition and image caption generation. Including attention improves the performance in these tasks [277–279] and can also lead to increases computational efficiency [280] and interpretability [281]. Another example of work in machine intelligence inspired by efficiency principle of the brain is neuromorphic computing. Neuromorphic computing aims to design brain inspired hardware. The goal is to improve the energy efficiency of hardware to improve the energy efficient of artificial intelligence, which is currently one of its main limitations [282].

Understanding how the brain makes efficient use of its limited resources can be relevant in other domains. For example, understanding efficiency can help us in our daily decisions. The neuroscientist Moran Cerf always picks the second item on the list of specials when eating out [283]. This algorithm may not select the best option, but saving on the costs of making the decision may make it preferable to considering all items on the menu. This approach is similar to the idea of "satisficing" [284], a heuristic according to which we should select the first option we encounter that is adequate, rather than consider all options and select the one with the highest utility. Depending on

the costs of making a decision and the marginal utility increase provided by finding better options, this algorithm may be preferable.

The principles of neural efficiency are also relevant for marketing. The number of different products sold in a typical supermarket increased from around 6 000 in the 1980s to around 30 000 today [285]. This increase was driven in part by horizontal segmentation [286], which is the idea that consumers should be proposed with multiple variations of a product in order to best fit their individual preferences. However, this increase in options can lead to choice overload [287]. Since considering options has a cost, increasing the options increases the mental load of the choice, which can lead consumers to prefer to not chose any product at all [288]. The supermarket chain Trader Joe's took another approach by limiting the number of different products they sell to around 2 000. This reduced the choice overload effect and can explain in part why the company is successful [289].

The neural efficiency principles may also help us understand our moral decisions. For example, consider an individual who wants to reduce their environmental footprint. Vegetarian and vegan meals may not always have a lower environmental footprint than meals with meat [290]. However, computing the environmental footprint of each meal is tedious, so avoiding animal products altogether can be considered an efficient heuristic to minimize one's environmental impact.

Finally, macroeconomists and policy makers can make use of behavior models that take into account how humans make decisions with limited resources. For example, given their limited cognitive resources, individuals will not adapt instantaneously and perfectly to new monetary policy [291], as they will need time to understand how the policies affect them and what their behavior should be. This "stickiness" should be taken into account when evaluating a given monetary policy [292]. Another important idea for policy making is that, as individuals have limited cognitive resources, they may be sensitive to the way the options are presented to them (e.g., which options are selected by default). The policy maker can then influence the choice of the individuals by changing the way in which the options are presented without changing the options themselves [293]. Although debated [294], it has been argued that since the options are not changed, the freedom to choose of individuals is the same [293]. This approach has been successfully used for example to increase participation in organ donation programs, by enrolling people by default instead of requiring them need to enroll [295].

We discussed how understanding that humans make efficient use of their limited resources in decision-making can help in different domains. In addition, considering efficiency altogether is also relevant. For example, since Trader Joe's sell less items, they have simplified their logistics. This can help explain why they make the highest revenue per store surface [289]. Another example can be found in hospital management, reducing the number of items in inventory can lead to reduced costs by saving time and space, reducing product waste, increasing clinician familiarity with products and improving work flow [296].

In the case of moral decision-making, we may consider the principles of efficiency differently. Moral decision-making is typically categorized in two ethical frameworks. Deontological ethics, which considers the morality of the action based on the action itself (e.g., lying is bad therefore we should not lie) and consequentialism, which considers the outcome of the actions (e.g., if lying saves lives, you should lie) [297]. The trolley problem is a thought experiment used to disentangle these two approaches [298]. In this dilemma, the decision maker is given the choice between an action that will result in the death of one individual (consequentialist choice) or doing nothing which will result in the death of five individuals (deontologist choice). Surprisingly, people tend to change their responses depending on the framing of the problem. Initial consequentialist become deontologist when they must push another person to stop the trolley instead of pulling a lever to deviate the trolley toward another person [299]. Therefore, it seems that individuals do not behave in a purely deontological or consequentialist way, but rather try to maximize the goodness of the consequences of their actions under the constraint of minimizing their deontological violations. This view is similar to the efficiency principle in decision-making discussed in this thesis, but instead of considering the cost of processing information, we consider a cost of violating ethical norms.

Considering efficiency is also relevant in macroeconomics. Current macroeconomic models rarely consider biophysical limitations such as energy, materials and waste [300]. In accordance, economics growth is usually considered a political goal. However, considering biophysical limitations suggests that economic growth is limited [301]. This has led to the development of post-growth frameworks which aim to decouple well-being from economic growth [302]. In other words, post-growth frameworks shift the focus from increasing the number of resources to efficiently using these resources. This is remarkably similar to the efficiency principle discussed in this thesis. We have observed that evolution likely did not simply pressure our ancestors to have larger

brain, but also to use them efficiently. Given that evolutionary biology and economics have historically inspired each other [303], one could expect this framework of efficiency to make its way into macroeconomics.

6.3. Closing remarks

Throughout this work, we have argued that the *when*, *what* and *how* of the brain are efficient. It is up to each of us to decide what to do with our efficient brain and to find the *purpose* of our efficiency. As J.R.R. Tolkien's Gandalf told Frodo, "All we have to decide is what to do with the time that is given us." [304]. In other words, our lifetimes are limited, and we are not able to do everything. The goal is not to do everything possible, but rather to *efficiently* live life according to our purpose, whatever we choose it to be.

A

APPENDIX: WEBER'S LAW: A MECHANISTIC FOUNDATION AFTER TWO CENTURIES

J. Brus*, J.A. Heng*, and R. Polanía, Weber's Law: A Mechanistic Foundation after Two Centuries. Cellpress: Trends in cognitive sciences (2019). doi:10.1016/j.tics.2019.09.001

A.1. Abstract

Weber's law appears to be a universal principle describing how we discriminate between physical magnitudes. However, this law remained purely descriptive for nearly two centuries. A study by Pardo-Vazquez et al. finally provides a mechanistic explanation, revealing how both accuracy and reaction-time performance lawfully emerge during sensory discrimination tasks.

A.2. Main

Weber's law (WL) [305] is one of the few psychophysical laws that are largely conserved across species and sensory modalities. WL states that, when comparing the magnitudes of two stimuli, our accuracy does not depend on the absolute intensity difference but instead on the ratio of the magnitudes of the compared stimuli. Crucially, this law results from empirical observations describing psychophysical performance, but ignores the temporal dynamics of the discrimination process. This lack of constraint has perhaps made it difficult to establish a biologically plausible mechanistic model of WL.

To solve this, Pardo-Vazquez et al. [47] designed an experiment that incorporates the time required for decision making. Specifically, they developed a sound intensity discrimination task performed by rats in which they varied from trial-to-trial the ratios and average magnitudes of the two sounds to be discriminated, Fig. A.1A. They observed that, although accuracy is constant for the same intensity ratio at different average magnitudes, the time required to make a decision does not follow this relationship. In particular, input

* These authors contributed equally to this work

stimuli with higher average magnitudes are more rapidly discriminated, Fig. A.1B,C, and their reaction-time distributions appear to be scale-invariant; that is, changes in average magnitude for a given intensity ratio are equivalent to a linear transformation in the units of time used to measure the decision times. Notably, although the importance of reaction times during magnitude discrimination was previously recognized [306, 307] the well-controlled experimental setup adopted in this study allowed, for the first time, researchers to demonstrate a tight relationship between scale-invariant reaction-time distributions and WL. Importantly, the authors rationalized that this apparent strict and joint requirement of accuracies and reaction times could provide a hint for establishing the mechanistic foundation of WL. To this end, the investigators relied on a general instantiation of a continuous Markov process model that allowed the dynamics of decision making to be flexibly captured.

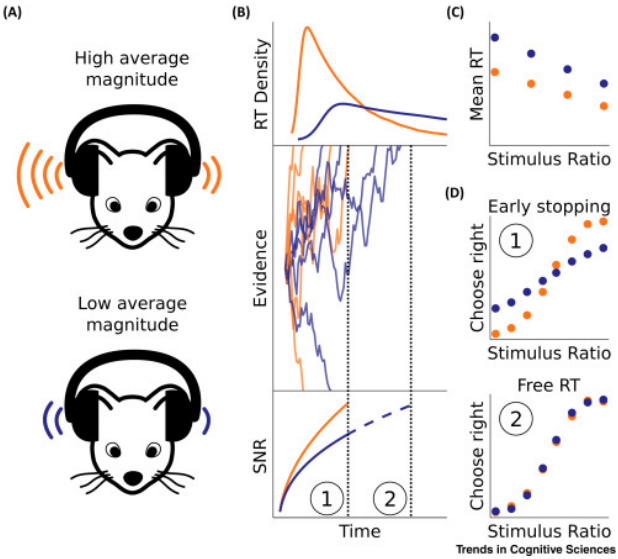


Fig. A.1: Dependencies of accuracy and reaction times on stimulus magnitude. **a)** Behavioral task: rats discriminated sounds at various ratios and average magnitudes, high/low magnitudes are in orange/blue. **b)** Rats accumulate evidence until reaching a decision threshold (middle). Higher magnitudes lead to faster but noisier accumulation, leading to scale-invariant reaction-time distributions (top). Signal-to-noise ratios at decision time are identical (bottom). **c)** Mean reaction times depend on sound ratio and average magnitude. **d)** If evidence accumulation is stopped early, trials with lower average magnitude have lower accuracy (1). However, for free reaction times, accuracy follows Weber's law (2). Abbreviations: RT, reaction time; SNR, signal to noise ratio.

First, they investigated the conditions necessary for a continuous Markov process to account for the psychophysical performance of WL and for the scale-invariant property of the reaction times. They found only one biologically plausible solution, which required a power-law encoding of stimulus magnitude as well as a linear relationship between the mean and variance of the sensory evidence. Notably, these requirements can be implemented by using populations of neurons with Poisson firing rates. In addition, a fixed decision threshold and the absence of decay in the accumulated evidence were the two remaining necessary conditions. As a result, the model belongs to the class of standard evidence-accumulation models which are commonly used in the decision-making literature [308]. These necessary properties not only allow a fundamental understanding of the decision mechanisms but are also suggested to be implemented by biological systems that instantiate decision processes.

The authors propose a parsimonious implementation of their mechanistic model, which was sufficient to capture the accuracy and reaction-time distributions of the animals even for data in a range of stimulus intensities that was not used to fit the model parameters. Interestingly, the new theory also generates a counterintuitive prediction about the breakdown of WL. Given that decision evidence evolves more slowly for lower sound intensities, the theory predicts that early stopping of stimulus presentation should lead to lower accuracies for quieter sounds, Fig. A.1B,D, which clearly violates WL. Strikingly, the results of experiments designed to incorporate this manipulation confirmed this counterintuitive prediction.

To provide evidence for the generality of their theory, Pardo-Vazquez et al. showed a similar dependency of accuracies and reaction-time distributions in humans on a similar sound discrimination task, and also in rodents in an odor mixture discrimination task. However, whether such a relationship holds for other modalities remains an open question. In particular, higher-order percepts (e.g., numerosity discrimination or reward-based decisions), which require integration of information in higher cortical areas, may follow distinct encoding rules from lower-order sensory systems such as those described in the study. An additional and intriguing result is that rats were unable to adapt the parameters of their decision-making process as a function of reward and motivation. The authors of the study investigated this by changing the rewards for correct trials depending on trial difficulty, and by presenting only the most difficult or easiest trials to the animals. Following principles of optimality, one would expect the animals to adapt their decision thresholds to maximize their reward rate [121]. Pardo-Vazquez et al. hypothesize that

this lack of adaptation could be due to the hardwired nature of neural systems dedicated to detecting interaural level differences in mammals [309]. Therefore, it could be argued that adaptation to reward distributions in this auditory system may require longer adaptation periods, perhaps via top-down influences of higher-order areas.

The clear exposition of the mechanisms underlying WL revealed by Pardo-Vazquez et al. generates new questions. For instance, it is unclear how the rigid relationships of accuracies and reaction times found in this study could account for contextual changes in the environment. In the case of the stimulus distributions used by Pardo-Vazquez et al. power-law encoding mechanisms have the convenient property of compressing physical stimulus intensity, allowing neuronal populations with a limited output range to represent a large spectrum of the physical world. Because the natural distributions of physical stimuli tend to follow a power law, such that lower-magnitude stimuli are more common than larger stimuli, power-law encoding allows better discrimination of frequently occurring stimuli, thus efficiently considering the allocation of limited encoding resources. Therefore, power-law encoding of a physical stimulus could be a product of computational principles such as efficient coding, which stems from the limited capacity of neural systems to represent information [89]. This predicts that encoding by sensory systems should adapt via experience and learning mechanisms if the stimulus distribution changes, and not only for early sensory perception but also for higher-order processing such as reward systems [74]. However, it is important to note that the sound intensities chosen by Pardo-Vazquez et al. were spaced logarithmically, which may have provided a similar distribution to naturally occurring stimuli. Thus, it would be interesting to place the animals in an environment with a different distribution of sound intensities (e.g., where higher intensities occur more frequently) to test whether WL and reaction-time distribution scale-invariance still holds.

As we finally move from a purely descriptive to a dynamic mechanistic explanation of WL, an interesting challenge for future research will be to understand how these mechanisms can be extended by incorporating learning and adaptation processes. Ultimately, organisms must constantly adapt to dynamic environments for survival. (Un)fortunately, because WL also applies to time perception [310], the additional decades of research to come will gradually be perceived as shorter and shorter.

Acknowledgments

This work was supported by a European Research Council (ERC) starting grant (ENTRAINER) to R.P. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 758604).

BIBLIOGRAPHY

[1] Marc Brysbaert. "How many words do we read per minute? A review and meta-analysis of reading rate". In: *Journal of Memory and Language* 109.April (2019), 104047. DOI: 10.1016/j.jml.2019.104047.

[2] Shuiyuan Yu, Chunshan Xu, and Haitao Liu. "Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation". In: (2018).

[3] G. K. Zipf. *Selected Studies of The Principle of Relative Frequency in Language*. Harvard University Press, 1932.

[4] Steven T Piantadosi. "Zipf's word frequency law in natural language: A critical review and future directions". In: *Psychonomic Bulletin and Review* 21.5 (2014), 1112. DOI: 10.3758/s13423-014-0585-6.

[5] Jasmeen Kanwal, Kenny Smith, Jennifer Culbertson, and Simon Kirby. "Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication". In: *Cognition* 165 (2017), 45. DOI: 10.1016/j.cognition.2017.05.001.

[6] Marcus E. Raichle and Debra A. Gusnard. *Appraising the brain's energy budget*. 2002. DOI: 10.1073/pnas.172399499.

[7] Ana Navarrete, Carel P Van Schaik, and Karin Isler. "Energetics and the evolution of human brain size". In: *Nature* 480.7375 (2011), 91. DOI: 10.1038/nature10629.

[8] Simon B Laughlin and Terrence J Sejnowski. *Communication in neuronal networks*. 2003. DOI: 10.1126/science.1089662.

[9] Christina Grimm, Sian N Duss, Mattia Privitera, Brandon R Munn, Stefan Frässle, Maria Chernysheva, Tommaso Patriarchi, Daniel Razansky, Nicole Wenderoth, James M Shine, Johannes Bohacek, and Valerio Zerbi. "Locus Coeruleus firing patterns selectively modulate brain activity and dynamics". In: *bioRxiv* (2022), 2022.08.29.505672. DOI: 10.1101/2022.08.29.505672.

[10] Gary Aston-Jones and Jonathan D. Cohen. *An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance*. 2005. DOI: 10.1146/annurev.neuro.28.061604.135709.

[11] Siddhartha Joshi, Yin Li, Rishi M. Kalwani, and Joshua I. Gold. "Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex". In: *Neuron* 89.1 (2016), 221. DOI: 10.1016/j.neuron.2015.11.028.

[12] Eckhard H. Hess and and James M. Polt. "Pupil Size in Relation to Mental Activity during Simple". In: *Science* 143.3611 (1964), 1190.

[13] Daniel Kahneman and Jackson Beatty. "Pupil diameter and load on memory". In: *Science* 154.3756 (1966), 1583. DOI: 10.1126/science.154.3756.1583.

[14] Marcus Grueschow, Birgit Kleim, and Christian C. Ruff. "Role of the locus coeruleus arousal system in cognitive control". In: *Journal of Neuroendocrinology* 32.12 (2020). DOI: 10.1111/jne.12890.

[15] Marco Pedrotti, Mohammad Ali Mirzaei, Adrien Tedesco, Jean Rémy Chardonnet, Frédéric Mérienne, Simone Benedetto, and Thierry Baccino. "Automatic Stress Classification With Pupil Diameter Analysis". In: *International Journal of Human-Computer Interaction* 30.3 (2014), 220. DOI: 10.1080/10447318.2013.848320.

[16] John O'Donnell, Douglas Zeppenfeld, Evan McConnell, Salvador Pena, and Maiken Nedergaard. *Norepinephrine: A neuromodulator that boosts the function of multiple cell types to optimize CNS performance.* 2012. DOI: 10.1007/s11064-012-0818-x.

[17] Hanna Hayat, Noa Regev, Noa Matosevich, Anna Sales, Elena Paredes-Rodriguez, Aaron J. Krom, Lottem Bergman, Yong Li, Marina Lavigne, Eric J. Kremer, Ofer Yizhar, Anthony E. Pickering, and Yuval Nir. "Locus coeruleus norepinephrine activity mediates sensory-evoked awakenings from sleep". In: *Science Advances* 6.15 (2020). DOI: 10.1126/SCIADV.AAZ4232.

[18] P. L. Madsen, J. F. Schmidt, G. Wildschiodtz, L. Friberg, S. Holm, S. Vorstrup, and N. A. Lassen. "Cerebral O2 metabolism and cerebral blood flow in humans during deep and rapid-eye-movement sleep". In: *Journal of Applied Physiology* 70.6 (1991), 2597. DOI: 10.1152/jappl.1991.70.6.2597.

[19] P. Maquet. "Sleep function(s) and cerebral metabolism". In: *Behavioural Brain Research* 69.1-2 (1995), 75. DOI: 10.1016/0166-4328(95)00017-N.

[20] Samuel J. Gershman and Taylor Burke. "Mental control of uncertainty". In: *Cognitive, Affective and Behavioral Neuroscience* 0123456789 (2022). DOI: 10.3758/s13415-022-01034-8.

[21] Annika Dix and Shu Chen Li. "Incentive motivation improves numerosity discrimination: Insights from pupillometry combined with drift-diffusion modelling". In: *Scientific Reports* 10.1 (2020). DOI: 10.1038/s41598-020-59415-3.

[22] James W. Bisley. *The neural basis of visual attention.* 2011. DOI: 10.1113/jphysiol.2010.192666.

[23] Isabella Rischall, Laura Hunter, Greg Jensen, and Jacqueline Gottlieb. "Inefficient prioritization of task-relevant attributes during instrumental information demand". In: *Nature Communications 2023 14:1* 14.1 (2023), 1. DOI: 10.1038/s41467-023-38821-x.

[24] H B Barlow and W A Rosenblith. "Possible principles underlying the transformations of sensory messages". In: *Sensory Communication.* Cambridge, MA: MIT Press, 1961.

[25] Deep Ganguli and Eero P. Simoncelli. "Neural and perceptual signatures of efficient sensory coding". In: (2016).

[26] Stuart Appelle. "Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals". In: *Psychological Bulletin* 78.4 (1972), 266. DOI: 10.1037/h0033117.

[27] Ari S. Benjamin, Ling Qi Zhang, Cheng Qiu, Alan A. Stocker, and Konrad P. Kording. "Efficient neural codes naturally emerge through gradient descent learning". In: *Nature Communications* 13.1 (2022), 1. DOI: 10.1038/s41467-022-35659-7.

[28] Jonathan Schaffner, Philippe N. Tobler, Todd A. Hare, and Rafael Polania. "Neural codes in early sensory areas maximize fitness". In: *bioRxiv* (2021), 2021.05.10.443388. DOI: 10.1101/2021.05.10.443388.

[29] Hans Berger. "Über das Elektrenkephalogramm des Menschen". In: *Archiv für Psychiatrie und Nervenkrankheiten* 87.1 (1929), 527. DOI: 10.1007/BF01797193.

[30] György Buzsáki. *Rhythms of the Brain.* Oxford University Press, 2006, 1. DOI: 10.1093/acprof:oso/9780195301069.001.0001.

[31] György Buzsáki, Nikos Logothetis, and Wolf Singer. *Scaling brain size, keeping timing: Evolutionary preservation of brain rhythms.* 2013. DOI: 10.1016/j.neuron.2013.10.002.

[32] Keith B. Doelling and M. Florencia Assaneo. "Neural oscillations are a start toward understanding brain activity rather than the end". In: *PLoS Biology* 19.5 (2021). DOI: 10.1371/journal.pbio.3001234.

[33] Matthew Chalk, Boris Gutkin, and Sophie Denève. "Neural oscillations as a signature of efficient coding in the presence of synaptic delays". In: *eLife* 5.2016JULY (2016). DOI: 10.7554/eLife.13824.

[34] Yvan Lengwiler. "The origins of expected utility theory". In: *Vinzenz Bronzin's Option Pricing Models: Exposition and Appraisal*. 2009, 535. DOI: 10.1007/978-3-540-85711-2_26.

[35] Herbert A. Simon. "A behavioral model of rational choice". In: *Quarterly Journal of Economics* 69.1 (1955), 99. DOI: 10.2307/1884852.

[36] Herbert A. Simon. "Models of Man: Social and Rational". In: (1957). DOI: 10.2307/3708691.

[37] Paul W. Glimcher. "Efficiently irrational: deciphering the riddle of human choice". In: *Trends in Cognitive Sciences* 26.8 (2022), 669. DOI: 10.1016/j.tics.2022.04.007.

[38] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. "Building machines that learn and think like people". In: *Behavioral and Brain Sciences* 40 (2017). DOI: 10.1017/S0140525X16001837.

[39] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. *Neuroscience-Inspired Artificial Intelligence*. 2017. DOI: 10.1016/j.neuron.2017.06.011.

[40] Anthony Zador, Blake Richards, Bence Ölveczky, Sean Escola, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, James DiCarlo, Surya Ganguli, Jeff Hawkins, Konrad Koerding, Alexei Koulakov, Yann LeCun, Timothy Lillicrap, Adam Marblestone, Bruno Olshausen, Alexandre Pouget, Cristina Savin, Terrence Sejnowski, Eero Simoncelli, Sara Solla, David Sussillo, Andreas S. Tolias, and Doris Tsao. "Toward Next-Generation Artificial Intelligence: Catalyzing the NeuroAI Revolution". In: *Nature Communications* (2023). DOI: 10.1038/s41467-023-37180-x.

[41] Elizabeth M. Brannon and Herbert S. Terrace. "Ordering of the numerosities 1 to 9 by monkeys". In: *Science* 282.5389 (1998), 746. DOI: 10.1126/science.282.5389.746.

[42] Hank Davis and Sheree Anne Bradford. "Counting Behavior by Rats in a Simulated Natural Environment". In: *Ethology* 73.4 (1986), 265. DOI: 10.1111/j.1439-0310.1986.tb00809.x.

[43] Damian Scarf, Harlene Hayne, and Michael Colombo. *Pigeons on par with primates in numerical competence*. 2011. DOI: 10.1126/science.1213357.

[44] Davide Potrich, Valeria Anna Sovrano, Gionata Stancher, and Giorgio Vallortigara. "Quantity discrimination by zebrafish (Danio rerio)". In: *Journal of Comparative Psychology* 129.4 (2015), 388. DOI: 10.1037/COM0000012.

[45] Scarlett R. Howard, Aurore Avarguès-Weber, Jair E. Garcia, Andrew D. Greentree, and Adrian G. Dyer. "Numerical ordering of zero in honey bees". In: *Science* 360.6393 (2018), 1124. DOI: 10.1126/science.aar4975.

[46] Khaled Nasr, Pooja Viswanathan, and Andreas Nieder. "Number detectors spontaneously emerge in a deep neural network designed for visual object recognition". In: *Science Advances* 5.5 (2019), eaav7903. DOI: 10.1126/sciadv.aav7903.

[47] Jose L. Pardo-Vazquez, Juan R. Castiñeiras-de Saa, Mafalda Valente, Iris Damião, Tiago Costa, M. Inês Vicente, André G. Mendonça, Zachary F. Mainen, and Alfonso Renart. "The mechanistic foundation of Weber's law". In: *Nature Neuroscience* 22.9 (2019), 1493. DOI: 10.1038/s41593-019-0439-7.

[48] Andreas Nieder. *The Adaptive Value of Numerical Competence*. 2020. DOI: 10.1016/j.tree.2020.02.009.

[49] Stanislas Dehaene. *The neural basis of the Weber-Fechner law: A logarithmic mental number line*. 2003. DOI: 10.1016/S1364-6613(03)00055-X.

[50] Samuel J. Cheyette and Steven T. Piantadosi. "A unified account of numerosity perception". In: *Nature Human Behaviour* 4.12 (2020), 1265. DOI: 10.1038/s41562-020-00946-0.

[51] Daniel Baldauf and Robert Desimone. "Neural mechanisms of object-based attention." In: *Science* 344.6182 (2014), 424. DOI: 10.1126/science.1247003.

[52] Thilo Womelsdorf, Jan Mathijs Schoffelen, Robert Oostenveld, Wolf Singer, Robert Desimone, Andreas K. Engel, and Pascal Fries. "Modulation of neuronal interactions through neuronal synchronization". In: *Science* 316.5831 (2007), 1609. DOI: 10.1126/SCIENCE.1139597.

[53] P. Fries, J. H. Reynolds, A. E. Rorie, and R. Desimone. "Modulation of oscillatory neuronal synchronization by selective visual attention". In: *Science* 291.5508 (2001), 1560. DOI: 10.1126/SCIENCE.1055465.

[54] Pascal Fries. "Rhythms For Cognition: Communication Through Coherence". In: *Neuron* 88.1 (2015), 220. DOI: 10.1016/J.NEURON.2015.09.034.

[55] Rafael Polanía, Michael A. Nitsche, and Christian C. Ruff. "Studying and modifying brain function with non-invasive brain stimulation". In: *Nature Neuroscience* 21.2 (2018), 174. DOI: 10.1038/s41593-017-0054-4.

[56] Valeriia Beliaeva, Iurii Savvateev, Valerio Zerbi, and Rafael Polania. "Toward integrative approaches to study the causal role of neural oscillations via transcranial electrical stimulation". In: *Nature Communications* 12.1 (2021), 2243. DOI: 10.1038/s41467-021-22468-7.

[57] Kevin R. Fontaine, David T. Redden, Chenxi Wang, Andrew O. Westfall, and David B. Allison. "Years of life lost due to obesity". In: *JAMA* 289.2 (2003), 187. DOI: 10.1001/JAMA.289.2.187.

[58] Maximilian Tremmel, Ulf G. Gerdtham, Peter M. Nilsson, and Sanjib Saha. *Economic burden of obesity: A systematic literature review*. 2017. DOI: 10.3390/ijerph14040435.

[59] Faidon Magkos, Inge Tetens, Susanne Gjedsted Bügel, Claus Felby, Simon Rønnow Schacht, James O. Hill, Eric Ravussin, and Arne Astrup. *The Environmental Foodprint of Obesity*. 2020. DOI: 10.1002/oby.22657.

[60] Nori Geary, Lori Asarian, Gwendolyn Graf, Susanna Gobbi, Philippe N Tobler, Jens F Rehfeld, and Brigitte Leeners. "Increased Meal Size but Reduced Meal-Stimulated Plasma Cholecystokinin Concentrations in Women With Obesity". In: *Endocrinology* 164.1 (2022). DOI: 10.1210/endocr/bqac192.

[61] Ivo Vlaev, Nick Chater, Neil Stewart, and Gordon D.A. Brown. "Does the brain calculate value?" In: *Trends in Cognitive Sciences* 15.11 (2011), 546. DOI: 10.1016/j.tics.2011.09.008.

[62] Benedetto De Martino and Aurelio Cortese. "Goals, usefulness and abstraction in value-based choice". In: *Trends in Cognitive Sciences* 27.1 (2023), 65. DOI: 10.1016/j.tics.2022.11.001.

[63] Todd A Hare, Colin F Camerer, and Antonio Rangel. "Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System". eng. In: *Science* 324.May (2009), 646. DOI: 10.1126/science.1168450.

[64] K. Carrie Armel, Aurelie Beaumel, and Antonio Rangel. "Biasing simple choices by manipulating relative visual attention". In: *Judgment and Decision Making* 3.5 (2008), 396. DOI: 10.1017/s1930297500000413.

[65] Ian Krajbich, Carrie Armel, and Antonio Rangel. "Visual fixations and the computation and comparison of value in simple choice". In: *Nature Neuroscience* 13.10 (2010), 1292. DOI: 10.1038/nn.2635.

[66] H. Barlow. "Redundancy reduction revisited". In: *Network: Computation in Neural Systems* 12.3 (2001), 241. DOI: 10.1080/net.12.3.241.253.

[67] Fred Attneave. "Some informational aspects of visual perception". In: *Psychological Review* 61.3 (1954), 183. DOI: 10.1037/h0054663.

[68] Jeremy E. Niven and Simon B. Laughlin. *Energy limitation as a selective pressure on the evolution of sensory systems.* 2008. DOI: 10.1242/jeb.017574.

[69] Tatyana O. Sharpee, Adam J. Calhoun, and Sreekanth H. Chalasani. *Information theory of adaptation in neurons, behavior, and mood.* 2014. DOI: 10.1016/j.conb.2013.11.007.

[70] Simon Laughlin. *A simple coding procedure enhances a neuron's information capacity.* 1981. DOI: 10.1515/znc-1981-9-1040.

[71] Deep Ganguli and Eero P. Simoncelli. "Efficient sensory encoding and Bayesian inference with heterogeneous neural populations". In: *Neural Computation* 26.10 (2014), 2103. DOI: 10.1162/NECO_a_00638.

[72] Xue Xin Wei and Alan A. Stocker. "A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts". In: *Nature Neuroscience* 18.10 (2015), 1509. DOI: 10.1038/nn.4105.

[73] Kenway Louie and Paul W Glimcher. "Efficient coding and the neural representation of value." In: *Annals of the New York Academy of Sciences* 1251 (2012), 13. DOI: 10.1111/j.1749-6632.2012.06496.x.

[74] Rafael Polanía, Michael Woodford, and Christian C. Ruff. "Efficient coding of subjective value". In: *Nature Neuroscience* 22.1 (2019), 134. DOI: 10.1038/s41593-018-0292-0.

[75] Aldo Rustichini, Katherine E. Conen, Xinying Cai, and Camillo Padoa-Schioppa. "Optimal coding and neuronal adaptation in economic decisions". In: *Nature Communications* 8.1 (2017), 1. DOI: 10.1038/s41467-017-01373-y.

[76] Susanne Schreiber, Christian K. Machens, Andreas V.M. Herz, and Simon B. Laughlin. "Energy-efficient coding with discrete stochastic events". In: *Neural Computation* 14.6 (2002), 1323. DOI: 10.1162/089976602753712963.

[77] Tatyana O. Sharpee. "Optimizing Neural Information Capacity through Discretization". In: *Neuron* 94.5 (2017), 954. DOI: 10.1016/j.neuron.2017.04.044.

[78] Michael N N. Shadlen and Daphna Shohamy. *Decision Making and Sequential Sampling from Memory*. 2016. DOI: 10.1016/j.neuron.2016.04.036.

[79] Donald A. Norman. "Toward a theory of memory and attention". In: *Psychological Review* 75.6 (1968), 522. DOI: 10.1037/h0026699.

[80] Elke U. Weber and Eric J. Johnson. *Mindful judgment and decision making*. 2009. DOI: 10.1146/annurev.psych.60.110707.163633.

[81] Neil Stewart, Nick Chater, and Gordon D.A. Brown. "Decision by sampling". In: *Cognitive Psychology* 53.1 (2006), 1. DOI: 10.1016/j.cogpsych.2005.10.003.

[82] Andreas Nieder and Stanislas Dehaene. *Representation of number in the brain*. 2009. DOI: 10.1146/annurev.neuro.051508.135550.

[83] Andreas Nieder and Earl K. Miller. "Coding of cognitive magnitude: compressed scaling of numerical information in the primate prefrontal cortex." In: *Neuron* 37.1 (2003), 149.

[84] Manuela Piazza, Philippe Pinel, Denis Le Bihan, and Stanislas Dehaene. "A Magnitude Code Common to Numerosities and Number Symbols in Human Intraparietal Cortex". In: *Neuron* 53.2 (2007), 293. DOI: 10.1016/j.neuron.2006.11.022.

[85] Mel Win Khaw, Ziang Li, and Michael Woodford. "Cognitive Imprecision and Small-Stakes Risk Aversion". In: *The Review of Economic Studies* (2020). DOI: 10.1093/restud/rdaa044.

[86] Michael Woodford. "Modeling Imprecision in Perception, Valuation, and Choice". In: *Annual Review of Economics* 12.1 (2020). DOI: 10.1146/annurev-economics-102819-040518.

[87] Brian Butterworth, C. R. Gallistel, and Giorgio Vallortigara. *Introduction: The origins of numerical abilities*. 2018.

[88] Jeroen Brus, Joseph A. Heng, and Rafael Polanía. "Weber's Law: A Mechanistic Foundation after Two Centuries". In: *Trends in Cognitive Sciences* 23.11 (2019), 906. DOI: 10.1016/J.TICS.2019.09.001.

[89] Xue Xin Wei and Alan A. Stocker. "Lawful relation between perceptual bias and discriminability". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.38 (2017), 10244. DOI: 10.1073/pnas.1619153114.

[90] Wiktor F. Młynarski and Ann M. Hermundstad. "Efficient and adaptive sensory codes". In: *Nature Neuroscience 2021 24:7* 24.7 (2021), 998. DOI: 10.1038/s41593-021-00846-0.

[91] Bertrand S. Clarke and Andrew R. Barron. "Jeffreys' prior is asymptotically least favorable under entropy risk". In: *Journal of Statistical Planning and Inference* 41.1 (1994), 37. DOI: 10.1016/0378-3758(94)90153-8.

[92] Il Memming Park and Jonathan W. Pillow. "Bayesian Efficient Coding". In: *bioRxiv* (2017), 1. DOI: 10.1101/178418.

[93] Mark D. McDonnell, Nigel G. Stocks, and Derek Abbott. "Optimal stimulus and noise distributions for information transmission via suprathreshold stochastic resonance". In: *Physical Review E* 75.6 (2007), 061105. DOI: 10.1103/PhysRevE.75.061105.

[94] Alexander P. Nikitin, Nigel G. Stocks, Robert P. Morse, and Mark D. McDonnell. "Neural population coding is optimized by discrete tuning curves". In: *Physical Review Letters* 103.13 (2009), 1. DOI: 10.1103/PhysRevLett.103.138101.

[95] Angelo Pirrone, Tom Stafford, and James A. R. Marshall. "When natural selection should optimize speed-accuracy trade-offs". In: *Frontiers in Neuroscience* 8 (2014), 73. DOI: 10.3389/fnins.2014.00073.

[96] Arman Abrahamyan, Laura Luz Silva, Steven C. Dakin, Matteo Carandini, and Justin L. Gardner. "Adaptable history biases in human perceptual decisions". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.25 (2016), E3548. DOI: 10.1073/pnas.1518786113.

[97] Waitsang Keung, Todd A. Hagen, and Robert C. Wilson. "Regulation of evidence accumulation by pupil-linked arousal processes". In: *Nature Human Behaviour* 3.6 (2019), 636. DOI: 10.1038/s41562-019-0551-4.

[98] Bharath Chandra Talluri, Anne E. Urai, Konstantinos Tsetsos, Marius Usher, and Tobias H. Donner. "Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence". In: *Current Biology* 28.19 (2018), 3128. DOI: 10.1016/j.cub.2018.07.052.

[99] Nils Kolling, Marco Wittmann, and Matthew F.S. Rushworth. "Multiple neural mechanisms of decision making and their competition under changing risk pressure". In: *Neuron* 81.5 (2014), 1190. DOI: 10.1016/j.neuron.2014.01.033.

[100] Klaas Enno Stephan, Will D. Penny, Jean Daunizeau, Rosalyn J. Moran, and Karl J. Friston. "Bayesian model selection for group studies". In: *NeuroImage* 46.4 (2009), 1004. DOI: 10.1016/j.neuroimage.2009.03.025.

[101] Gilles Dutilh and Jörg Rieskamp. "Comparing perceptual and preferential decision making". In: *Psychonomic Bulletin and Review* 23.3 (2016), 723. DOI: 10.3758/s13423-015-0941-1.

[102] Rafael Polanía, Ian Krajbich, Marcus Grueschow, and Christian C. Ruff. "Neural Oscillations and Synchronization Differentially Support Evidence Accumulation in Perceptual and Value-Based Decision Making". In: *Neuron* 82.3 (2014), 709. DOI: 10.1016/j.neuron.2014.03.014.

[103] Marcus Grueschow, Rafael Polania, Todd A. Hare, and Christian C. Ruff. "Automatic versus Choice-Dependent Value Representations in the Human Brain". In: *Neuron* 85.4 (2015), 874. DOI: 10.1016/j.neuron.2014.12.054.

[104] K. Hawkes, J. F. O'Connell, N. G. Blurton Jones, H. Alvarez, and E. L. Charnov. "Grandmothering, menopause, and the evolution of human life histories". In: *Proceedings of the National Academy of Sciences of the United States of America* 95.3 (1998), 1336. DOI: 10.1073/pnas.95.3.1336.

[105] James V Stone. "Principles of neural information processing". In: *Cognitive Systems Monographs* 27 (2018). DOI: 10.1007/978-3-319-20113-9_1.

[106] Emilio Salinas. "How behavioral constraints may determine optimal sensory representations". In: *PLoS Biology* 4.12 (2006), 2383. DOI: 10.1371/journal.pbio.0040387.

[107] Roger Ratcliff and Jeffrey N. Rouder. "Modeling Response Times for Two-Choice Decisions". In: *Psychological Science* 9.5 (1998), 347. DOI: 10.1111/1467-9280.00067.

[108] Xiao Jing Wang. "Probabilistic Decision Making by Slow Reverberation in Cortical Circuits". In: *Neuron* 36.5 (2002), 955. DOI: 10.1016/S0896-6273(02)01092-9.

[109] A. Aldo Faisal, Luc P.J. Selen, and Daniel M. Wolpert. *Noise in the nervous system*. 2008. DOI: 10.1038/nrn2258.

[110] Konstantinos Tsetsos, Rani Moran, James Moreland, Nick Chater, Marius Usher, and Christopher Summerfield. "Economic irrationality is optimal during noisy decision making". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.11 (2016), 3102. DOI: 10.1073/pnas.1519157113.

[111] Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. "Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex". In: *Neuron* 92.2 (2016), 530. DOI: 10.1016/j.neuron.2016.09.038.

[112] Go Ashida and Masayoshi Kubo. "Suprathreshold stochastic resonance induced by ion channel fluctuation". In: *Physica D: Nonlinear Phenomena* 239.6 (2010), 327. DOI: 10.1016/J.PHYSD.2009.12.002.

[113] Brett A. Schmerl and Mark D. McDonnell. "Channel-noise-induced stochastic facilitation in an auditory brainstem neuron model". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 88.5 (2013), 1. DOI: 10.1103/PhysRevE.88.052722.

[114] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D. Cox, and James J. DiCarlo. "Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations". In: *bioRxiv* (2020), 2020.06.16.154542. DOI: 10.1101/2020.06.16.154542.

[115] Charles Findling and Valentin Wyart. "Computation noise promotes cognitive resilience to adverse conditions during decision-making". In: *bioRxiv* (2020), 2020.06.10.145300. DOI: 10.1101/2020.06.10.145300.

[116] Douglas D. Garrett, Natasa Kovacevic, Anthony R. McIntosh, and Cheryl L. Grady. "The importance of being variable". In: *Journal of Neuroscience* 31.12 (2011), 4496. DOI: 10.1523/JNEUROSCI.5641-10.2011.

[117] Kenneth W. Latimer, Jacob L. Yates, Miriam L.R. Meister, Alexander C. Huk, and Jonathan W. Pillow. "Single-trial spike trains in parietal cortex reveal discrete steps during decision-making". In: *Science* 349.6244 (2015), 184. DOI: 10.1126/SCIENCE.AAA4056.

[118] David M. Zoltowski, Kenneth W. Latimer, Jacob L. Yates, Alexander C. Huk, and Jonathan W. Pillow. "Discrete Stepping and Nonlinear Ramping Dynamics Underlie Spiking Responses of LIP Neurons during Decision-Making". In: *Neuron* 102.6 (2019), 1249. DOI: 10.1016/j.neuron.2019.04.031.

[119] Rahul Bhui and Samuel J. Gershman. "Decision by sampling implements efficient coding of psychoeconomic functions". In: *Psychological Review* 125.6 (2018), 985. DOI: 10.1037/rev0000123.

[120] Katherine L. Buchanan, Jennifer L. Grindstaff, and Vladimir V. Pravosudov. "Condition dependence, developmental plasticity, and cognition: Implications for ecology and evolution". In: *Trends in Ecology and Evolution* 28.5 (2013), 290. DOI: 10.1016/j.tree.2013.02.004.

[121] Satohiro Tajima, Jan Drugowitsch, and Alexandre Pouget. "Optimal policy for value-based decision-making". In: *Nature Communications 2016 7:1* 7.1 (2016), 1. DOI: 10.1038/ncomms12400.

[122] Rafael Polanía, Marius Moisa, Alexander Opitz, Marcus Grueschow, and Christian C. Ruff. "The precision of value-based choices depends causally on fronto-parietal phase coupling". In: *Nature Communications* 6.1 (2015), 1. DOI: 10.1038/ncomms9090.

[123] Aaron M. Bornstein, Mel W. Khaw, Daphna Shohamy, and Nathaniel D. Daw. "Reminders of past choices bias decisions for reward in humans". In: *Nature Communications* 8.1 (2017), 15958. DOI: 10.1038/ncomms15958.

[124] Sebastian Gluth, Tobias Sommer, Jörg Rieskamp, and Christian Büchel. "Effective Connectivity between Hippocampus and Ventromedial Prefrontal Cortex Controls Preferential Choices from Memory". In: *Neuron* 86.4 (2015), 1078. DOI: 10.1016/j.neuron.2015.04.023.

[125] Aman B. Saleem, E. Mika Diamanti, Julien Fournier, Kenneth D. Harris, and Matteo Carandini. "Coherent encoding of subjective spatial position in visual cortex and hippocampus". In: *Nature 2018 562:7725* 562.7725 (2018), 124. DOI: 10.1038/s41586-018-0516-1.

[126] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco, CA: W.H. Freeman, 1982.

[127] Ahmad T. Qamar, R. James Cotton, Ryan G. George, Jeffrey M. Beck, Eugenia Prezhdo, Allison Laudano, Andreas S. Tolias, and Wei Ji Ma. "Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization". In: *Proceedings of the National Academy of Sciences of the United States of America* 110.50 (2013), 20332. DOI: 10.1073/pnas.1219756110.

[128] Justin L. Gardner. "Optimality and heuristics in perceptual neuroscience". In: *Nature Neuroscience* (2019), 1. DOI: 10.1038/s41593-019-0340-4.

[129] Bingni W Brunton, Matthew M Botvinick, and Carlos D Brody. "Rats and humans can optimally accumulate evidence for decision-making." In: *Science (New York, N.Y.)* 340.6128 (2013), 95. DOI: 10.1126/science.1233912.

[130] Alan A. Stocker and Eero P. Simoncelli. "Noise characteristics and prior expectations in human visual speed perception". In: *Nature Neuroscience* 9.4 (2006), 578. DOI: 10.1038/nn1669.

[131] Ariel Zylberberg, Daniel M. Wolpert, and Michael N. Shadlen. "Counterfactual Reasoning Underlies the Learning of Priors in Decision Making". In: *Neuron* 99.5 (2018), 1083. DOI: 10.1016/j.neuron.2018.07.035.

[132] Ronald van den Berg, Marcus Lindskog, Leo Poom, and Anders Winman. "Recent is more: A negative time-order effect in nonsymbolic numerical judgment". In: *Journal of Experimental Psychology: Human Perception and Performance* 43.6 (2017), 1084. DOI: 10.1037/xhp0000387.

[133] Véronique Izard and Stanislas Dehaene. "Calibrating the mental number line". In: *Cognition* 106.3 (2008), 1221. DOI: 10.1016/J.COGNITION.2007.06.004.

[134] Aki Vehtari, Andrew Gelman, and Jonah Gabry. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and Computing* 27.5 (2017), 1413. DOI: 10.1007/s11222-016-9696-4.

[135] N. G. Stocks, D. Allingham, and R. P. Morse. "the Application of Suprathreshold Stochastic Resonance To Cochlear Implant Coding". In: *Fluctuation and Noise Letters* 02.03 (2002), L169. DOI: 10.1142/s0219477502000774.

[136] Giovanni Anobile, Roberto Arrighi, Elisa Castaldi, and David C. Burr. "A Sensorimotor Numerosity System". In: *Trends in Cognitive Sciences* 25.1 (2021), 24. DOI: 10.1016/J.TICS.2020.10.009.

[137] Susannah K. Revkin, Manuela Piazza, Véronique Izard, Laurent Cohen, and Stanislas Dehaene. "Does subitizing reflect numerical estimation?" In: *Psychological Science* 19.6 (2008), 607. DOI: 10.1111/j.1467-9280.2008.02130.x.

[138] Giovanni Anobile, Guido Marco Cicchini, and David C. Burr. "Linear mapping of numbers onto space requires attention". In: *Cognition* 122.3 (2012), 454.

[139] John Whalen, C. R. Gallistel, and Rochel Gelman. "Nonverbal Counting in Humans: The Psychophysics of Number Representation". In: *Psychological Science* 10.2 (1999), 130. DOI: 10.1111/1467-9280.00120.

[140] Matthew Inglis and Camilla Gilmore. "Sampling from the mental number line: How are approximate number system representations formed?" In: *Cognition* 129.1 (2013), 63. DOI: 10.1016/J.COGNITION.2013.06.003.

[141] Ana Navarrete, Carel P. van Schaik, and Karin Isler. "Energetics and the evolution of human brain size". In: *Nature* 480.7375 (2011), 91. DOI: 10.1038/nature10629.

[142] Rahul Bhui, Lucy Lai, and Samuel J. Gershman. *Resource-rational decision making*. 2021. DOI: 10.1016/j.cobeha.2021.02.015.

[143] Roger Ratcliff. "A theory of memory retrieval". In: *Psychological Review* 85.2 (1978), 59. DOI: 10.1037/0033-295X.85.2.59.

[144] Pedro A. Ortega and Daniel A. Braun. "Thermodynamics as a theory of decision-making with information-processing costs". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 469.2153 (2013). DOI: 10.1098/RSPA.2012.0683.

[145] Pedro A. Ortega and Alan A. Stocker. "Human decision-making under limited time". In: *Advances in Neural Information Processing Systems* 29 (2016).

[146] Karl Friston. "The free-energy principle: a rough guide to the brain?" In: *Trends in Cognitive Sciences* 13.7 (2009), 293. DOI: 10.1016/J.TICS.2009.04.005.

[147] D. H. Wolpert. "Complex Engineering Systems, chapter Information theory-the bridge connecting bounded rational game theory and statistical physics". In: Perseus Books, 2004.

[148] Eckhard Limpert, Werner A. Stahel, and Markus Abbt. "Log-normal distributions across the sciences: keys and clues". In: *BioScience* 51.5 (2001), 341. DOI: 10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2.

[149] Steven T. Piantadosi. "A rational analysis of the approximate number system". In: *Psychonomic Bulletin and Review* 23.3 (2016), 877. DOI: 10.3758/S13423-015-0963-8.

[150] Stanislas Dehaene and Jacques Mehler. "Cross-linguistic regularities in the frequency of number words". In: *Cognition* 43.1 (1992), 1. DOI: 10.1016/0010-0277(92)90030-L.

[151] M. M. Taylor, P. H. Lindsay, and S. M. Forbes. "Quantification of shared capacity processing in auditory and visual discrimination". In: *Acta Psychologica* 27 (1967), 223. DOI: 10.1016/0001-6918(67)99000-2.

[152] Thomas M Cover. *Elements of information theory*. John Wiley \& Sons, 1999.

[153] Jeffrey S. Bowers and Colin J. Davis. "Bayesian just-so stories in psychology and neuroscience." In: *Psychological Bulletin* 138.3 (2012), 389. DOI: 10.1037/a0026450.

[154] Alan A. Stocker and Eero P. Simoncelli. "Noise characteristics and prior expectations in human visual speed perception." In: *Nature Neuroscience* 9.4 (2006), 578. DOI: 10.1038/nn1669.

[155] Ahna R. Girshick, Michael S. Landy, and Eero P. Simoncelli. "Cardinal rules: visual orientation perception reflects knowledge of environmental statistics". In: *Nature Neuroscience* 14.7 (2011), 926. DOI: 10.1038/nn.2831.

[156] Christoph Teufel and Paul C. Fletcher. "Forms of prediction in the nervous system". In: *Nature Reviews Neuroscience* (2020), 1. DOI: 10.1038/s41583-020-0275-5.

[157] Joseph A. Heng, Michael Woodford, and Rafael Polania. "Efficient sampling and noisy decisions". In: *eLife* 9 (2020). DOI: 10.7554/eLife.54962.

[158] Alberto Testolin and James L. McClelland. "Do estimates of numerosity really adhere to Weber's law? A reexamination of two case studies". In: *Psychonomic Bulletin and Review* 28.1 (2021), 158. DOI: 10.3758/s13423-020-01801-z.

[159] Cary Frydman and Lawrence J. Jin. "Efficient Coding and Risky Choice". In: *The Quarterly Journal of Economics* 137.1 (2021), 161. DOI: 10.1093/QJE/QJAB031.

[160] Michael Woodford. "Prospect Theory as Efficient Perceptual Distortion". In: *American Economic Review* 102.3 (2012), 41. DOI: 10.1257/aer.102.3.41.

[161] Nikola Grujic, Jeroen Brus, Denis Burdakov, and Rafael Polania. "Rational inattention in mice". In: *Science Advances* 8.9 (2022), 8935. DOI: 10.1126/SCIADV.ABJ8935.

[162] Katharine Mullen, David Ardia, David Gil, Donald Windover, and James Cline. "{DEoptim}: An {R} Package for Global Optimization by {D}ifferential {E}volution". In: *Journal of Statistical Software* 40.6 (2011), 1. DOI: 10.18637/jss.v040.i06.

[163] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2017.

[164] Stefan Treue and Julio C. Martínez Trujillo. "Feature-based attention influences motion processing gain in macaque visual cortex". In: *Nature* 399.6736 (1999), 575. DOI: 10.1038/21176.

[165] Theodore P. Zanto, Michael T. Rubens, Arul Thangavel, and Adam Gazzaley. "Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory". In: *Nature Neuroscience* 14.5 (2011), 656. DOI: 10.1038/nn.2773.

[166] Marco Bedini and Daniel Baldauf. "Structure, function and connectivity fingerprints of the frontal eye field versus the inferior frontal junction: A comprehensive comparison". In: *European Journal of Neuroscience* 54.4 (2021), 5462. DOI: 10.1111/ejn.15393.

[167] Ayelet Nina Landau and Pascal Fries. "Attention Samples Stimuli Rhythmically". In: *Current Biology* 22.11 (2012), 1000. DOI: 10.1016/J.CUB.2012.03.054.

[168] Laura Dugué, Mariel Roberts, and Marisa Carrasco. "Attention Reorients Periodically". In: *Current Biology* 26.12 (2016), 1595. DOI: 10.1016/J.CUB.2016.04.046.

[169] Ian C. Fiebelkorn, Yuri B. Saalmann, and Sabine Kastner. "Rhythmic Sampling within and between Objects despite Sustained Attention at a Cued Location". In: *Current Biology* 23.24 (2013), 2553. DOI: 10.1016/J.CUB.2013.10.063.

[170] Randolph F. Helfrich, Ian C. Fiebelkorn, Sara M. Szczepanski, Jack J. Lin, Josef Parvizi, Robert T. Knight, and Sabine Kastner. "Neural Mechanisms of Sustained Attention Are Rhythmic". In: *Neuron* 99.4 (2018), 854. DOI: 10.1016/j.neuron.2018.07.032.

[171] Geoffrey Brookshire. "Putative rhythms in attentional switching can be explained by aperiodic temporal structure". In: *Nature Human Behaviour* (2022), 1. DOI: 10.1038/s41562-022-01364-0.

[172] Robert M.G. Reinhart and John A. Nguyen. "Working memory revived in older adults by synchronizing rhythmic brain circuits". In: *Nature Neuroscience* 22.5 (2019), 820. DOI: 10.1038/s41593-019-0371-x.

[173] Nina Wolinski, Nicholas R. Cooper, Paul Sauseng, and Vincenzo Romei. "The speed of parietal theta frequency drives visuospatial working memory capacity". In: *PLOS Biology* 16.3 (2018), e2005348. DOI: 10.1371/JOURNAL.PBIO.2005348.

[174] Luke Johnson, Ivan Alekseichuk, Jordan Krieg, Alex Doyle, Ying Yu, Jerrold Vitek, Matthew Johnson, and Alexander Opitz. "Dose-dependent effects of transcranial alternating current stimulation on spike timing in awake nonhuman primates". In: *Science Advances* 6.36 (2020), eaaz2747. DOI: 10.1126/sciadv.aaz2747.

[175] Pedro G. Vieira, Matthew R. Krause, and Christopher C. Pack. "tACS entrains neural activity while somatosensory input is blocked". In: *PLoS Biology* 18.10 (2020). DOI: 10.1371/journal.pbio.3000834.

[176] Mohsin M Ali, Kristin K Sellers, Flavio Fröhlich, and F. Frohlich. "Transcranial alternating current stimulation modulates large-scale cortical network activity by network resonance." In: *The Journal of Neuroscience* 33.27 (2013), 11262. DOI: 10.1523/JNEUROSCI.5867-12.2013.

[177] Min Fang Kuo, Rafael Polanía, and Michael Nitsche. "Physiology of transcranial direct and alternating current stimulation". In: *Transcranial Direct Current Stimulation in Neuropsychiatric Disorders: Clinical Principles and Management* (2016), 29. DOI: 10.1007/978-3-319-33967-2_3.

[178] Sam Ling, Taosheng Liu, and Marisa Carrasco. "How spatial and feature-based attention affect the gain and tuning of population responses". In: *Vision Research* 49.10 (2009), 1194. DOI: 10.1016/J.VISRES.2008.05.025.

[179] M G Philiastides, R Auksztulewicz, H R Heekeren, and F Blankenburg. "Causal role of dorsolateral prefrontal cortex in human perceptual decision making". eng. In: *Curr Biol* 21.11 (2011), 980. DOI: 10.1016/j.cub.2011.04.034.

[180] Boateng Asamoah, Ahmad Khatoun, and Myles Mc Laughlin. "tACS motor system effects can be caused by transcutaneous stimulation of peripheral nerves". In: *Nature Communications* (2019). DOI: 10.1038/s41467-018-08183-w.

[181] Christopher A Hill, Shinsuke Suzuki, Rafael Polania, Marius Moisa, John P O'Doherty, and Christian C Ruff. "A causal account of the brain network computations underlying strategic social behavior." In: *Nature neuroscience* (2017). DOI: 10.1038/nn.4602.

[182] Valeriia Beliaeva and Rafael Polania. "Can low-intensity tACS genuinely entrain neural activity in vivo?" In: *Brain Stimulation* 13.6 (2020), 1796. DOI: 10.1016/j.brs.2020.10.002.

[183] Zeynep M. Saygin, David E. Osher, Kami Koldewyn, Gretchen Reynolds, John D.E. Gabrieli, and Rebecca R. Saxe. "Anatomical connectivity patterns predict face selectivity in the fusiform gyrus". In: *Nature Neuroscience* 15.2 (2012), 321. DOI: 10.1038/nn.3001.

[184] Alison Mary, Jacques Dayan, Giovanni Leone, Charlotte Postel, Florence Fraisse, Carine Malle, Thomas Vallée, Carine Klein-Peschanski, Fausto Viader, Vincent de la Sayette, Denis Peschanski, Francis Eustache, and Pierre Gagnepain. "Resilience after trauma: The role of memory suppression". In: *Science* 367.6479 (2020). DOI: 10.1126/science.aay8477.

[185] Shrey Grover, Wen Wen, Vighnesh Viswanathan, Christopher T. Gill, and Robert M.G. Reinhart. "Long-lasting, dissociable improvements in working memory and long-term memory in older adults with repetitive neuromodulation". In: *Nature Neuroscience* 25.9 (2022), 1237. DOI: 10.1038/s41593-022-01132-3.

[186] Javid Sadr and Pawan Sinha. "Object recognition and Random Image Structure Evolution". In: *Cognitive Science* 28.2 (2004), 259. DOI: 10.1016/j.cogsci.2003.09.003.

[187] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan Mathijs Schoffelen. "FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data". In: *Computational intelligence and neuroscience* 2011 (2011). DOI: 10.1155/2011/156869.

[188] Martin Vinck, Robert Oostenveld, Marijn Van Wingerden, Franscesco Battaglia, and Cyriel M.A. Pennartz. "An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias". In: *NeuroImage* 55.4 (2011), 1548. DOI: 10.1016/J.NEUROIMAGE.2011.01.055.

[189] Barry D. Van Veen, Wim Van Drongelen, Moshe Yuchtman, and Akifumi Suzuki. "Localization of brain electrical activity via linearly constrained minimum variance spatial filtering". In: *IEEE transactions on bio-medical engineering* 44.9 (1997), 867. DOI: 10.1109/10.623056.

[190] Maria Ida Iacono, Esra Neufeld, Esther Akinnagbe, Kelsey Bower, Johanna Wolf, Ioannis Vogiatzis Oikonomidis, Deepika Sharma, Bryn Lloyd, Bertram J. Wilm, Michael Wyss, Klaas P. Pruessmann, Andras Jakab, Nikos Makris, Ethan D. Cohen, Niels Kuster, Wolfgang Kainz, and Leonardo M. Angelone. "MIDA: A multimodal imaging-based detailed anatomical model of the human head and neck". In: *PLoS ONE* 10.4 (2015). DOI: 10.1371/journal.pone.0124126.

[191] PA Hasgall, F Di Gennaro, C Baumgartner, E Neufeld, B Lloyd, MC Gosselin, D Payne, A Klingenböck, and N Kuster. "IT'IS Database for thermal and electromagnetic parameters of biological tissues". In: *It'Is* (2018). DOI: 10.13099/VIP21000-04-1.

[192] Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R. Laird, Peter T. Fox, Simon B. Eickhoff, Chunshui Yu, and Tianzi Jiang. "The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture". In: *Cerebral Cortex* 26.8 (2016), 3508. DOI: 10.1093/cercor/bhw157.

[193] Chad A. Bossetti, Merrill J. Birdno, and Warren M. Grill. "Analysis of the quasi-static approximation for calculating potentials generated by neural stimulation". In: *Journal of Neural Engineering* 5.1 (2008), 44. DOI: 10.1088/1741-2560/5/1/005.

[194] Ralf Engbert and Reinhold Kliegl. "Microsaccades uncover the orientation of covert attention". In: *Vision Research* 43.9 (2003), 1035. DOI: 10.1016/S0042-6989(03)00084-1.

[195] Deborah W. Tang, Lesley K. Fellows, and Alain Dagher. "Behavioral and Neural Valuation of Foods Is Driven by Implicit Knowledge of Caloric Content". In: *Psychological Science* 25.12 (2014), 2168. DOI: 10.1177/0956797614552081.

[196] Shinsuke Suzuki, Logan Cross, and John P. O'Doherty. "Elucidating the underlying components of food valuation in the human orbitofrontal cortex". In: *Nature Neuroscience* 20.12 (2017), 1780. DOI: 10.1038/s41593-017-0008-x.

[197] Alexandra G. DiFeliceantonio, Géraldine Coppin, Lionel Rigoux, Sharmili Edwin Thanarajah, Alain Dagher, Marc Tittgemeyer, and Dana M. Small. "Supra-Additive Effects of Combining Fat and Carbohydrate on Food Reward". In: *Cell Metabolism* 28.1 (2018), 33. DOI: 10.1016/j.cmet.2018.05.018.

[198] Todd A Hare, Jonathan Malmaud, and Antonio Rangel. "Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice." In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 31.30 (2011), 11077. DOI: 10.1523/JNEUROSCI.6383-10.2011.

[199] Silvia U. Maier, Anjali Raja Beharelle, Rafael Polanía, Christian C. Ruff, and Todd A. Hare. "Dissociable mechanisms govern when and how strongly reward attributes affect decisions". In: *Nature Human Behaviour* (2020), 1. DOI: 10.1038/s41562-020-0893-y.

[200] Hilke Plassmann, John O'Doherty, Baba Shiv, and Antonio Rangel. "Marketing actions can modulate neural representations of experienced pleasantness". In: *Proceedings of the National Academy of Sciences of the United States of America* 105.3 (2008), 1050. DOI: 10.1073/pnas.0706929105.

[201] Charles Spence, Carmel A. Levitan, Maya U. Shankar, and Massimiliano Zampini. *Does food color influence taste and flavor perception in humans?* 2010. DOI: 10.1007/s12078-010-9067-z.

[202] Katsunori Okajima and Charles Spence. "Effects of Visual Food Texture on Taste Perception". In: *i-Perception* 2.8 (2011), 966. DOI: 10.1068/ic966.

[203] Douglas G. Lee and Todd A. Hare. "Evidence Accumulates for Individual Attributes During Value-Based Decisions". In: *Decision* (2022), 2021.08.05.455296. DOI: 10.1037/dec0000190.

[204] Valentina Marques da Rosa, Charles Spence, and Leandro Miletto Tonetto. "Influences of visual attributes of food packaging on consumer preference and associations with taste and healthiness". In: *International Journal of Consumer Studies* 43.2 (2019), 210. DOI: 10.1111/ijcs.12500.

[205] Kathryn M. Wall, Michael C. Farruggia, Emily E. Perszyk, Arsene Kanyamibwa, Sophie Fromm, Xue S. Davis, Jelle R. Dalenberg, Alexandra G. DiFeliceantonio, and Dana M. Small. "No evidence for an association between obesity and milkshake liking". In: *International Journal of Obesity* 44.8 (2020), 1668. DOI: 10.1038/s41366-020-0583-x.

[206] Emily E. Perszyk, Zach Hutelin, Jessica Trinh, Arsene Kanyamibwa, Sophie Fromm, Xue S. Davis, Kathryn M. Wall, Kyle D. Flack, Alexandra G. Difeliceantonio, and Dana M. Small. "Fat and carbohydrate interact to potentiate food reward in healthy weight but not in overweight or obesity". In: *Nutrients* 13.4 (2021). DOI: 10.3390/nu13041203.

[207] E. H. Castellanos, E. Charboneau, M. S. Dietrich, S. Park, B. P. Bradley, K. Mogg, and R. L. Cowan. "Obese adults have visual attention bias for food cue images: Evidence for altered reward system function". In: *International Journal of Obesity* 33.9 (2009), 1063. DOI: 10.1038/ijo.2009.138.

[208] Ilse M.T. Nijs, Peter Muris, Anja S. Euser, and Ingmar H.A. Franken. "Differences in attention to food and food intake between overweight/obese and normal-weight females under conditions of hunger and satiety". In: *Appetite* 54.2 (2010), 243. DOI: 10.1016/j.appet.2009.11.004.

[209] Reiko Graham, Alison Hoover, Natalie A. Ceballos, and Oleg Komogortsev. "Body mass index moderates gaze orienting biases and pupil diameter to high and low calorie food images". In: *Appetite* 56.3 (2011), 577. DOI: 10.1016/j.appet.2011.01.029.

[210] Jessica Werthmann, Anne Roefs, Chantal Nederkoorn, Karin Mogg, Brendan P. Bradley, and Anita Jansen. "Can(not) Take my Eyes off it: Attention Bias for Food in Overweight Participants". In: *Health Psychology* 30.5 (2011), 561. DOI: 10.1037/a0024291.

[211] Kathrin Schag, Martin Teufel, Florian Junne, Hubert Preissl, Martin Hautzinger, Stephan Zipfel, and Katrin Elisabeth Giel. "Impulsivity in Binge Eating Disorder: Food Cues Elicit Increased Reward Responses and Disinhibition". In: *PLoS ONE* 8.10 (2013), e76542. DOI: 10.1371/journal.pone.0076542.

[212] Katy J. Doolan, Gavin Breslin, Donncha Hanna, Kate Murphy, and Alison M. Gallagher. "Visual attention to food cues in obesity: An eye-tracking study". In: *Obesity* 22.12 (2014), 2501. DOI: 10.1002/oby.20884.

[213] Jessica Werthmann, Anita C.E. Vreugdenhil, Anita Jansen, Chantal Nederkoorn, Ghislaine Schyns, and Anne Roefs. "Food Through the Child's Eye: An Eye-Tracking Study on Attentional Bias for Food in Healthy-Weight Children and Children With Obesity". In: *Health Psychology* 34.12 (2015), 1123. DOI: 10.1037/hea0000225.

[214] Kelsey E. Hagan, Ahmed Alasmar, Alexis Exum, Bernadette Chinn, and Kelsie T. Forbush. *A systematic review and meta-analysis of attentional bias toward food in individuals with overweight and obesity.* 2020. DOI: 10.1016/j.appet.2020.104710.

[215] Chantal Julia and Serge Hercberg. "Nutri-Score: evidence of the effective-ness of the French front-of-pack nutrition label". In: *Ernahrungs Umschau* 64.12 (2017), 181. DOI: 10.4455/eu.2017.048.

[216] Roger Ratcliff and Gail McKoon. *The diffusion decision model: Theory and data for two-choice decision tasks.* 2008. DOI: 10.1162/neco.2008.12-06-420.

[217] Stephanie M. Smith and Ian Krajbich. "Gaze Amplifies Value in Decision Making". In: *Psychological Science* 30.1 (2019), 116. DOI: 10.1177/0956797618810521.

[218] Sarah Lichtenstein and Paul Slovic. "Reversals of preference between bids and choices in gambling decisions". In: *Journal of Experimental Psychology* 89.1 (1971), 46. DOI: 10.1037/H0031207.

[219] Betty E. Kim, Darryl Seligman, and Joseph W. Kable. "Preference reversals in decision making under risk are accompanied by changes in attention to different attributes". In: *Frontiers in Neuroscience* 6.JULY (2012), 1. DOI: 10.3389/FNINS.2012.00109.

[220] Carlos Alós-Ferrer, Alexander Jaudas, and Alexander Ritschel. "Attentional shifts and preference reversals: An eye-tracking study". In: *Judgment and Decision Making* 16.1 (2021), 57.

[221] Douglas Lee and Jean Daunizeau. "Choosing what we like vs liking what we choose: How choice-induced preference change might actually be instrumental to decision-making". In: *PLoS ONE* 15.5 (2020), e0231081. DOI: 10.1371/journal.pone.0231081.

[222] Fergus M. Clydesdale. "Color as a factor in food choice". In: *Critical Reviews in Food Science and Nutrition* 33.1 (1993), 83. DOI: 10.1080/10408399309527614.

[223] H. Alexander Chen, Iris B. Hovens, Xue S. Davis, Zach Hutelin, Kathryn M. Wall, and Dana M. Small. "Identification of a novel link between adiposity and visuospatial perception". In: *Obesity* 31.2 (2023), 423.

[224] Suzanne Higgs and Maartje S Spetter. "Cognitive Control of Eating: the Role of Memory in Appetite and Weight Gain". In: (2018). DOI: 10.1007/s13679-018-0296-9.

[225] Fania C M Dassen, Katrijn Houben, Vanessa Allom, and Anita Jansen. "Self-regulation and obesity: the role of executive function and delay discounting in the prediction of weight loss". In: *Journal of Behavioral Medicine* 41 (2018). DOI: 10.1007/s10865-018-9940-9.

[226] Rachael Gwinn, Andrew B. Leber, and Ian Krajbich. "The spillover effects of attentional learning on value-based choice". In: *Cognition* 182 (2019), 294. DOI: 10.1016/j.cognition.2018.10.012.

[227] Howard R. Moskowitz. "Relative Importance of Perceptual Factors to Consumer Acceptance: Linear vs Quadratic Analysis". In: *Journal of Food Science* 46.1 (1981), 244. DOI: 10.1111/j.1365-2621.1981.tb14573.x.

[228] Howard R. Moskowitz, Robert A. Kluter, Judith Westerling, and Harry L. Jacobs. "Sugar Sweetness and Pleasantness: Evidence for Different Psychological Laws". In: *Science* 184.4136 (1974), 583. DOI: 10.1126/SCIENCE.184.4136.583.

[229] Osayanmon Wellington Osawe, Gianluca Grilli, and John Curtis. "Examining food preferences in the face of environmental pressures". In: *Journal of Agriculture and Food Research* 11 (2023), 100476. DOI: 10.1016/j.jafr.2022.100476.

[230] Tegan Cruwys, Kirsten E. Bevelander, and Roel C.J. Hermans. "Social modeling of eating: A review of when and why social influence affects food intake and choice". In: *Appetite* 86 (2015), 3. DOI: 10.1016/j.appet.2014.08.035.

[231] Erica Van De Waal, Christèle Borgeaud, and Andrew Whiten. "Potent social learning and conformity shape a wild primate's foraging decisions". In: *Science* 340.6131 (2013), 483. DOI: 10.1126/science.1232769.

[232] Sudeep Bhatia and Neil Stewart. "Naturalistic multiattribute choice". In: *Cognition* 179 (2018), 71. DOI: 10.1016/J.COGNITION.2018.05.025.

[233] Jana B. Jarecki and Jörg Rieskamp. "Comparing attribute-based and memory-based preferential choice". In: *Decision* 49.1 (2022), 65. DOI: 10.1007/s40622-021-00302-9.

[234] Ivan E. De Araujo, Mark Schatzker, and Dana M. Small. "Rethinking Food Reward". In: *https://doi.org/10.1146/annurev-psych-122216-011643* 71 (2020), 1. DOI: 10.1146/ANNUREV-PSYCH-122216-011643.

[235] Shinsuke Suzuki. "Constructing value signals for food rewards: determinants and the integration". In: *Current Opinion in Behavioral Sciences* 46 (2022). DOI: 10.1016/J.COBEHA.2022.101178.

[236] Mickael Camus, Neil Halelamien, Hilke Plassmann, Shinsuke Shimojo, John O'Doherty, Colin Camerer, and Antonio Rangel. "Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex decreases valuations during food choices". In: *European Journal of Neuroscience* 30.10 (2009), 1980. DOI: 10.1111/j.1460-9568.2009.06991.x.

[237] Sebastian Gluth, Nadja Kern, Maria Kortmann, and Cécile L Vitali. "Value-based attention but not divisive normalization influences decisions with multiple alternatives". In: *Nature Human Behaviour* (). DOI: 10.1038/s41562-020-0822-0.

[238] Douglas G. Lee and Keith J. Holyoak. "Coherence Shifts in Attribute Evaluations". In: *Decision* 8.4 (2021), 257. DOI: 10.1037/dec0000151.

[239] P. Read Montague, Raymond J. Dolan, Karl J. Friston, and Peter Dayan. "Computational psychiatry". In: *Trends in Cognitive Sciences* 16.1 (2012), 72. DOI: 10.1016/J.TICS.2011.11.018.

[240] Tanja V.E. Kral, Reneé H. Moore, Jesse Chittams, Elizabeth Jones, Lauren O'Malley, and Jennifer O. Fisher. "Identifying behavioral phenotypes for childhood obesity". In: *Appetite* 127 (2018), 87. DOI: 10.1016/J.APPET.2018.04.021.

[241] Leonard Kozarzewski, Lukas Maurer, Anja Mähler, Joachim Spranger, and Martin Weygandt. *Computational approaches to predicting treatment response to obesity using neuroimaging*. Vol. 23. 4. Springer US, 2022, 773. DOI: 10.1007/s11154-021-09701-w.

[242] Alex Zvoleff. *glcm: Calculate Textures from Grey-Level Co-Occurrence Matrices (GLCMs)*. 2020.

[243] Jeroen Ooms. *magick: Advanced Graphics and Image-Processing in R*. 2021.

[244] David Hasler and Sabine E. Suesstrunk. "Measuring colorfulness in natural images". In: *Human Vision and Electronic Imaging VIII*. Vol. 5007. SPIE, 2003, 87. DOI: 10.1117/12.477378.

[245] David H. Brainard. "The Psychophysics Toolbox". In: *Spatial Vision* 10.4 (1997), 433. DOI: 10.1163/156856897X00357.

[246] D. G. Pelli. "The VideoToolbox software for visual psychophysics: Transforming numbers into movies". In: *Spatial Vision* 10 (1997), 437.

[247] Mario Kleiner, David Brainard, Denis Pelli, Allen Ingling, Richard Murray, and Christopher Broussard. *What's new in Psychtoolbox-3 ? Perception 36 ECVP Abstract Supplement*. 2007.

[248] Diederick C. Niehorster, Richard Andersson, and Marcus Nyström. "Titta: A toolbox for creating PsychToolbox and Psychopy experiments with Tobii eye trackers". In: *Behavior Research Methods* 52.5 (2020), 1970.

[249] John Kruschke. *Doing Bayesian Data Analysis (Second Edition)*. Academic Press, 2015.

[250] Andrew Gelman, Jennifer Hill, and Masanao Yajima. "Why We (Usually) Don't Have to Worry About Multiple Comparisons". In: *Journal of Research on Educational Effectiveness* 5.2 (2012), 189. DOI: 10.1080/19345747.2011.618213.

[251] M. Plummer. "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. 2003.

[252] Matthew J. Denwood. "runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS". In: *Journal of Statistical Software* 71 (2016), 1. DOI: 10.18637/JSS.V071.I09.

[253] Martyn Plummer. *rjags: Bayesian Graphical Models using MCMC*. 2019.

[254] Ron Ofri. "Retina". In: *Slatter's Fundamentals of Veterinary Ophthalmology* (2008), 285. DOI: 10.1016/B978-072160561-6.50018-6.

[255] Stefan R. Schweinberger, Christoph Casper, Nadine Hauthal, Jürgen M. Kaufmann, Hideki Kawahara, Nadine Kloth, David M.C. Robertson, Adrian P. Simpson, and Romi Zäske. "Auditory Adaptation in Voice Perception". In: *Current Biology* 18.9 (2008), 684. DOI: 10.1016/j.cub.2008.04.015.

[256] Daniel Kaping, Paul Duhamel, and Michael A. Webster. "Adaptation to natural facial categories". In: *Journal of Vision* 2.10 (2002), 357. DOI: 10.1167/2.10.128.

[257] David Burr and John Ross. "A Visual Sense of Number". In: *Current Biology* 18.6 (2008), 425. DOI: 10.1016/j.cub.2008.02.052.

[258] Steuart Henderson Britt. "How Weber's law can be applied to marketing". In: *Business Horizons* 18.1 (1975), 21. DOI: 10.1016/0007-6813(75)90004-X.

[259] Miguel Barretto Garcia, Marcus Grueschow, Rafael Polania, Michael Woodford, and Christian Ruff. "Individual risk attitudes arise from noise in neurocognitive magnitude representations". In: *Research Square* (2022). DOI: 10.21203/rs.3.rs-1989825/v1.

[260] Andreas Nieder. "Coding of abstract quantity by 'number neurons' of the primate brain". In: *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* 199.1 (2013), 1. DOI: 10.1007/s00359-012-0763-9.

[261] Vikram S. Chib, Antonio Rangel, Shinsuke Shimojo, and John P. O'Doherty. "Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex". In: *Journal of Neuroscience* 29.39 (2009), 12315. DOI: 10.1523/JNEUROSCI.2575-09.2009.

[262] Arthur Prat-Carrabin and Michael Woodford. "Efficient coding of numbers explains decision bias and noise". In: *Nature Human Behaviour 2022 6:8* 6.8 (2022), 1142. DOI: 10.1038/s41562-022-01352-4.

[263] Wei Ji Ma and Michael Woodford. "Multiple conceptions of resource rationality". In: *Behavioral and Brain Sciences* 43 (2020), e15. DOI: 10.1017/S0140525X19001754.

[264] Christopher J. Bates, Chris R. Sims, and Robert A. Jacobs. "The importance of constraints on constraints". In: *Behavioral and Brain Sciences* 43 (2020), e3. DOI: 10.1017/S0140525X19001572.

[265] Susannah K. Revkin, Manuela Piazza, Véronique Izard, Laurent Cohen, and Stanislas Dehaene. "Does subitizing reflect numerical estimation?" In: *Psychological Science* 19.6 (2008), 607. DOI: 10.1111/j.1467-9280.2008.02130.x.

[266] H. Choo and S. L. Franconeri. "Enumeration of small collections violates Weber's law". In: *Psychonomic Bulletin and Review* 21.1 (2014), 93. DOI: 10.3758/s13423-013-0474-4.

[267] Azadeh Hajihosseini and Cendri A Hutcherson. "Alpha oscillations and event-related potentials reflect distinct dynamics of attribute construction and evidence accumulation in dietary decision making". In: *eLife* (2021), 1. DOI: 10.7554/eLife.60874.

[268] Earl K. Miller, Mikael Lundqvist, and André M. Bastos. "Working Memory 2.0". In: *Neuron* 100.2 (2018), 463. DOI: 10.1016/J.NEURON.2018.09.023.

[269] Giovanna Aiello, Debora Ledergeber, Tena Dubcek, Lennart Stieglitz, Christian Baumann, Rafael Polanìa, and Lukas Imbach. "Functional network dynamics between the anterior thalamus and the cortex in deep brain stimulation for epilepsy". In: *Brain* (2023). DOI: 10.1093/BRAIN/AWAD211.

[270] Ashwini Oswal, Peter Brown, and Vladimir Litvak. "Synchronized neural oscillations and the pathophysiology of Parkinson's disease". In: *Current Opinion in Neurology* 26.6 (2013), 662. DOI: 10.1097/WCO.0000000000000034.

[271] Erol Başar. "Brain oscillations in neuropsychiatric disease". In: *Dialogues in Clinical Neuroscience* 15.3 (2013), 291. DOI: 10.31887/dcns.2013.15.3/ebasar.

[272] Ryan T. Canolty and Robert T. Knight. "The functional role of cross-frequency coupling". In: *Trends in Cognitive Sciences* 14.11 (2010), 506. DOI: 10.1016/j.tics.2010.09.001.

[273] Lindsey Smith Taillie, Marcela Reyes, M. Arantxa Colchero, Barry Popkin, and Camila Corvalán. "An evaluation of Chile's Law of Food Labeling and Advertising on sugar-sweetened beverage purchases from 2015 to 2017: A before-and-after study". In: *PLOS Medicine* 17.2 (2020), e1003015. DOI: 10.1371/JOURNAL.PMED.1003015.

[274] Peter R. Murphy, Joachim Vandekerckhove, and Sander Nieuwenhuis. "Pupil-Linked Arousal Determines Variability in Perceptual Decision Making". In: *PLOS Computational Biology* 10.9 (2014), e1003854. DOI: 10.1371/JOURNAL.PCBI.1003854.

[275] Martin J. Dahl, Mara Mather, and Markus Werkle-Bergner. "Noradrenergic modulation of rhythmic neural activity shapes selective attention". In: *Trends in Cognitive Sciences* 26.1 (2022), 38. DOI: 10.1016/j.tics.2021.10.009.

[276] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. "An Attentive Survey of Attention Models". In: *ACM Transactions on Intelligent Systems and Technology* 12.5 (2021), 1. DOI: 10.1145/3465055.

[277] Feng Wang and David M. J. Tax. "Survey on the attention based RNN model and its applications in computer vision". In: (2016).

[278] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015), 1. DOI: arXiv:1409.0473.

[279] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. "Recurrent models of visual attention". In: *Advances in Neural Information Processing Systems* 3.January (2014), 2204.

[280] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems* 2017-Decem (2017), 5999.

[281] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. "Bias in BIOS: A case study of semantic representation bias in a high-stakes setting". In: *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (2019), 120. DOI: 10.1145/3287560.3287572.

[282] Md Sakibul Islam, Sharif Noor Zisad, Ah Lian Kor, and Md Hasibul Hasan. "Sustainability of Machine Learning Models: An Energy Consumption Centric Evaluation". In: *3rd International Conference on Electrical, Computer and Communication Engineering, ECCE 2023* (2023), 1. DOI: 10.1109/ECCE57851.2023.10101532.

[283] Chris Weller. *A neuroscientist explains why he always picks the 2nd menu item on a list of specials*. 2017. DOI: https://www.businessinsider.com/neuroscientist-decision-making-hack-restaurants-2017-7.

[284] Herbert A. Simon. "Rational Choice and the Structure of the Environment". In: *Psychological Review.* (1956). DOI: 10.1037/h0042769.

[285] Susan M. Broniarczyk and Wayne D. Hoyer. "Retail Assortment: More \neq Better". In: *Springer Books* (2010), 271. DOI: 10.1007/978-3-540-72003-4_17.

[286] Howard R. Moskowitz, Barry E. Jacobs, and Neil Lazar. "Product response segmentation and the analysis of individual differences in liking". In: *Journal of Food Quality* 8.2-3 (1985), 169. DOI: 10.1111/j.1745-4557.1985.tb00844.x.

[287] Alexander Chernev, Ulf Böckenholt, and Joseph Goodman. "Choice overload: A conceptual review and meta-analysis". In: *Journal of Consumer Psychology* 25.2 (2015), 333. DOI: 10.1016/J.JCPS.2014.08.002.

[288] Sheena S. Iyengar and Mark R. Lepper. "When choice is demotivating: Can one desire too much of a good thing?" In: *Journal of Personality and Social Psychology* 79.6 (2000), 995. DOI: 10.1037/0022-3514.79.6.995.

[289] Stephen J. Dubner. *Should America Be Run by ... Trader Joe's?* 2018. DOI: https://freakonomics.com/podcast/should-america-be-run-by-trader-joes/.

[290] J. Poore and T. Nemecek. "Reducing food's environmental impacts through producers and consumers". In: *Science* 360.6392 (2018), 987. DOI: 10.1126/science.aaq0216.

[291] Michael Woodford. "Macroeconomic Analysis Without the Rational Expectations Hypothesis". In: *https://doi.org/10.1146/annurev-economics-080511-110857* 5 (2013), 303. DOI: 10.1146/ANNUREV-ECONOMICS-080511-110857.

[292] Mariana García-Schmidt and Michael Woodford. "Are low interest rates deflationary? A paradox of perfect-foresight analysis†". In: *American Economic Review* 109.1 (2019), 86. DOI: 10.1257/aer.20170110.

[293] Cass R. Sunstein. "Nudging: A Very Short Guide". In: *Journal of Consumer Policy* 37.4 (2014), 583. DOI: 10.1007/s10603-014-9273-1.

[294] Evan Selinger and Kyle Whyte. "Is There a Right Way to Nudge? The Practice and Ethics of Choice Architecture". In: *Sociology Compass* 5.10 (2011), 923. DOI: 10.1111/j.1751-9020.2011.00413.x.

[295] Shai Davidai, Thomas Gilovich, and Lee D. Ross. "The meaning of default options for potential organ donors". In: *Proceedings of the National Academy of Sciences of the United States of America* 109.38 (2012), 15201. DOI: 10.1073/pnas.1211695109.

[296] P. Delatore, M. Bourque, J. Taylor, and N. Ferko. "The value of SKU reduction and standardization initiatives within a hospital system". In: *Value in Health* 19.3 (2016), A293. DOI: 10.1016/j.jval.2016.03.843.

[297] Subrata Chakrabarty and A. Erin Bass. "Comparing Virtue, Consequentialist, and Deontological Ethics-Based Corporate Social Responsibility: Mitigating Microfinance Risk in Institutional Voids". In: *Journal of Business Ethics* 126.3 (2015), 487. DOI: 10.1007/s10551-013-1963-0.

[298] Phillipa Foot. "The Problem of Abortion and the Doctrine of the Double Effect* Phillipa Foot Before reading the article by Foot, this short discussion of the doctrine of the double effect is useful." In: *Oxford Review* 5 (1967).

[299] J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen. "An fMRI investigation of emotional engagement in moral judgment". In: *Science* 293.5537 (2001), 2105. DOI: DOI:10.1126/science.1062872.

[300] Pierre Jacques, Louis Delannoy, Baptiste Andrieu, Devrim Yilmaz, Hervé Jeanmart, and Antoine Godin. "Assessing the economic consequences of an energy transition through a biophysical stock-flow consistent model". In: *Ecological Economics* 209 (2023), 107832. DOI: 10.1016/J.ECOLECON.2023.107832.

[301] H Donella Meadows, Dennis Meadows, Jorgen Randers, and William W Behrens III. *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind.* Potomac Associates – Universe Books, 1972.

[302] Stefano Bartolini and Francesco Sarracino. "Happier and Sustainable. Possibilities for a post-growth society". In: *Munich Personal RePEc Archive* 108309 (2021).

[303] Johannes Martens. "Economics and evolutionary biology : an overview of their (recent) interactions". In: (2020). DOI: hal-03089648.

[304] J. R. R. Tolkien. *The Fellowship of the Ring.* George Allen & Unwin, 1954.

[305] E. H. Weber. *De Pulsu, resorptione, auditu et tactu. Annotationes Anatomicae et Physiologicae.* C.F. Koehler, 1834.

[306] Stephen W. Link. *The Wave Theory of Difference and Similarity.* Routledge, 2020. DOI: 10.1201/9780429054709.

[307] Patrick Simen, Ksenia Vlasov, and Samantha Papadakis. "Scale (In)variance in a unified diffusion model of decision making and timing". In: *Psychological Review* 123.2 (2016), 151. DOI: 10.1037/rev0000014.

[308] Roger Ratcliff, Philip L. Smith, Scott D. Brown, and Gail McKoon. *Diffusion Decision Model: Current Issues and History.* 2016. DOI: 10.1016/j.tics.2016.01.007.

[309] Benedikt Grothe, Michael Pecka, and David McAlpine. *Mechanisms of sound localization in mammals.* 2010. DOI: 10.1152/physrev.00026.2009.

[310] Mehrdad Jazayeri and Michael N. Shadlen. "Temporal context calibrates interval timing". In: *Nature Neuroscience* 13.8 (2010), 1020. DOI: 10.1038/nn.2590.

Joseph Heng

PHD STUDENT IN NEUROSCIENCE

Born March 1993 | French and American| joseph.heng@hest.ethz.ch
linkedin.com/in/joseph-heng-86608457

Experience

ETHZ – Decision Neuroscience Lab | Doctoral researcher **2018 – Present**

- Studying brain and behavior with a combination of behavioral modeling, neuroimaging (EEG, fMRI and eye-tracking) and neurostimulation (non-invasive transcranial stimulation)
- Research focus: numerosity perception, non-spatial attention, dietary decision-making in obesity and sweetness perception
- Teaching assistant for the course "Bayesian Data Analysis and Models of Behavior"
- Supervising bachelor and master students

World Health Organization | Intern **2017**

- Cleaned data for the Mental Health Atlas 2017

UCSD – Schwartz Center for Computational Neuroscience | Master thesis **2016 – 2017**

- Developed a toolbox for EEG data analysis

Education

ETHZ, PhD in Neuroscience **2018 – Present**

EPFL, Master in Life Sciences and Technology – Neuroscience and Neuroengineering **2014 – 2017**

Minor in computational neuroscience

EPFL, Bachelor in Life Sciences and Technology **2011 – 2014**

Awarded EPFL Excellence Fellowship

Exchange year at Polytechnique Montréal

Publications

R. Martínez-Cancino, **J. A. Heng**, A. Delorme, K. Kreutz-Delgado, R. C. Sotero, S. Makeig (2018) Measuring transient phase-amplitude coupling using local mutual information. ***Neuroimage.***

J. Brus*, **J. A. Heng***, Rafael Polanía. (2019) Weber's Law: A Mechanistic Foundation after Two Centuries. ***Trends in Cognitive Sciences.*** (*Shared first author)

J. A. Heng, M. Woodford, R. Polanía (2020) Efficient sampling and noisy decisions. ***eLife.***