The Bayesian Origin of the Probability Weighting Function in Human Representation of Probabilities

Xin Tong[1]* **Thi Thu Uyen Hoang**[1] **Xue-Xin Wei**[2] **Michael Hahn**[1]†
[1]Saarland University, Saarland Informatics Campus [2]The University of Texas at Austin

Abstract

Humans systematically misrepresent probability in a stereotyped inverse-S pattern. It has been documented for decades, but its origin remains unexplained. We propose a Bayesian encoding–decoding account in which probabilities are represented by noisy internal signals and decoded by Bayes-risk minimization. For bounded probability stimuli, we show that distortion decomposes into boundary regression, likelihood repulsion, and prior attraction, yielding a key prediction: the classic inverse-S-shaped weighting pattern implies a U-shaped allocation of encoding precision with greater sensitivity near 0 and 1. Across judgment of relative frequency, lottery pricing, and risky choice, this U-shape is recovered from data without imposing any functional form on the encoding, and our framework outperforms deterministic weighting functions, bounded log-odds models, uniform-encoding Bayesian accounts, and matched efficient-coding models on held-out data. In a new dot probability estimation experiment with bimodal stimulus statistics, the recovered prior tracks the new distribution while the recovered encoding remains U-shaped. Together, these results identify the inverse-S-shaped probability weighting function as the joint product of a stable U-shaped encoding and a flexible prior, integrated by optimal Bayesian decoding.[3]

1 Introduction

Humans' representation of probability is distorted. A 10% chance of an event may feel higher than 10%, while a 90% chance may feel lower than 90%. Prospect Theory [16] formalized this overweighting of small probabilities and underweighting of large ones as an inverse-S-shaped probability weighting function (Figure 1A). Probability weighting function represents one of the most robust empirical findings in behavioral decision research [23], and is central for explaining anomalies such as the Allais paradox [1].

A fundamental question remains: what is the origin of the probability weighting function? Classical parametric forms [22, 29] describe its shape but offer no explanatory account. A growing body of work points to a more fundamental answer: that the distortion reflects how the brain encodes probability, rather than how decisions are computed from it. Yet the literature offers strikingly different proposals for what that representation looks like. Some accounts attribute the distortion to a log-odds transformation of noisy estimates [30, 17]. Others locate it in regression away from response boundaries, with encoding itself unbiased [7, 3]. Still others derive it from efficient coding, in which the encoding precisely matches the prior distribution of probabilities in the environment [10]. Each proposal implies a different picture of how the brain represents probability, and each has been validated on a different task. Whether they are competing explanations, complementary mechanisms, or special cases of a common principle has not been resolved.

*Contact: xintongcontact@gmail.com
†Contact: mhahn@lst.uni-saarland.de
[3]Code: https://github.com/aklywtx/bayesian-origin-probability-distortion

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In this paper, we address this question by building on the Bayesian encoding–decoding framework of Hahn and Wei [12]. We extend its analysis of bias distortion to the bounded domain $[0, 1]$, deriving how the previously identified mechanisms—boundary regression, likelihood repulsion, and prior attraction—are modulated near either endpoint. This extension yields a key prediction: the classic inverse-S weighting pattern implies a U-shaped allocation of encoding precision, with sensitivity peaking near 0 and 1. We confirm this prediction empirically across three behavioral paradigms (judgment of relative frequency, lottery pricing, and risky choice), recovering the U-shape from data without imposing any functional form on the encoding.

Beyond identifying this shared mechanism, we establish formal connections among regression-based, log-odds, bounded log-odds, and efficient-coding accounts, clarifying their distinct assumptions about encoding, priors, and decoding. We compare these accounts empirically, finding that held-out predictions generally favor the freely fitted encoding–prior model over more constrained alternatives. Finally, in a new dot-estimation experiment with bimodal stimulus statistics, the recovered prior tracks the stimulus distribution while the encoding remains U-shaped. Together, these findings characterize probability distortion as the joint product of a stable U-shaped encoding and a flexible prior that tracks the environment.

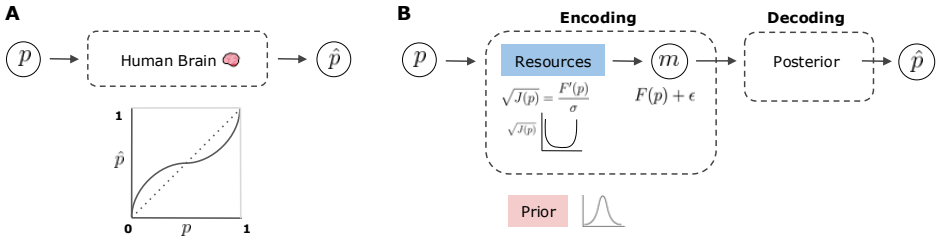


Figure 1: Distorted probability perception and our Bayesian model that explains such perceptual distortion. (A) Human decisions systematically distort probabilities, producing the inverse S-shaped probability weighting function central to Prospect Theory. (B) We propose that the inverse-S weighting pattern in (A) emerges from optimal decoding of a U-shaped allocation of encoding resources, recovered empirically from behavior. The Bayesian encoding–decoding framework: probabilities p are encoded noisily, combined with a prior, and decoded into perceived probabilities $\hat{p}$.

2 Background and Relevant Work

Existing accounts of probability distortion fall into two broad families. The first family, represented by a broad body of work in economics and psychology, models the inverse-S pattern through a deterministic parametric function, the *probability weighting function* $w(p)$. Classical examples include the inverse-S shaped forms in Prospect Theory [25], the Prelec function [22], and the Linear-in-Log-Odds (LILO) model [29]. As an illustration, the LILO model assumes that under the log-odds transform, the perceived probability $\widehat{p}$ is linear in the true probability p. These models successfully capture behavioral regularities such as overweighting of small probabilities and underweighting of large ones, but they are primarily descriptive.

The question of where this shape comes from has motivated a second family of work, which focuses on the idea that probabilities are encoded imprecisely in the brain, with distortion arising as a systematic consequence of noisy internal representations. Within this view, accounts diverge on where the noise enters and what structure the encoding has. Regression-based accounts locate the bias at the response stage, attributing distortion to regression away from the scale boundaries while leaving encoding itself unbiased [7, 3]. Log-odds based accounts locate the bias at the encoding stage, assuming that internal representations are approximately linear in log-odds space [17, 18]. Efficient-coding accounts further constrain the encoding to match the prior distribution of probabilities in the environment [10], while related work has examined optimal inference under noisy encoding more broadly [15, 5]. These models have only been evaluated against limited behavioral data, and their differing predictions have not been compared on common ground, either theoretically or empirically.

Among encoding-based accounts, the Bounded Log-Odds (BLO) model [30] stands out as the most quantitatively developed. It has been validated across multiple behavioral paradigms and consistently

outperforms fixed probability weighting functions, making it the natural benchmark against which any new account must be measured. Because our empirical comparisons (Section 4) treat BLO as the primary reference point, we review its mechanism in detail here. BLO assumes that probability p is first mapped to log-odds and truncated to a bounded interval $[\Delta^-, \Delta^+]$ (Eq. 1). The clipped value is linearly mapped to an encoding $\Lambda(p)$ on an interval $[-\Psi, \Psi]$ (Eq. 2), then combined with an anchor Λ_0 via a weight that depends on a free parameter $\kappa > 0$ (Eq. 3). This encoding is subject to Gaussian noise and decoded by applying the inverse log-odds function (Eq. 4).

$$\lambda(p) = \log \tfrac{p}{1-p}, \qquad \Gamma(\lambda) = \min\left(\max(\lambda, \Delta^-), \Delta^+\right) \tag{1}$$

$$\Lambda(p) = \tfrac{\Psi}{(\Delta_+ - \Delta_-)/2} \left(\Gamma(\lambda(p)) - \tfrac{\Delta_- + \Delta_+}{2}\right) \tag{2}$$

$$\hat{\Lambda}_\omega(p) = \omega_p \cdot \Lambda(p) + (1 - \omega_p) \cdot \Lambda_0, \qquad \omega_p = \tfrac{1}{1+\kappa V(p)}, \qquad V(p) \propto p(1-p), \tag{3}$$

$$\hat{\pi}(p) = \lambda^{-1}(\hat{\Lambda}_\omega(p) + \epsilon_\lambda), \quad \epsilon_\lambda \sim \mathcal{N}(0, \sigma_\lambda^2). \tag{4}$$

3 A general Bayesian framework for perceived probability

3.1 Encoding–decoding framework

Across the encoding-based accounts reviewed in Section 2, models differ in their choice of encoding function and prior, but share a common structure: probabilities are encoded into a noisy internal signal, and a perceived probability is decoded from this signal. We make this shared structure explicit by formalizing it as a single Bayesian encoding–decoding framework. Within this framework, the existing accounts, including BLO, regression-based, log-odds-based, and efficient-coding models, correspond to particular choices of two ingredients: the encoding function F and the prior P_{prior}. This unification will allow us in Section 3.2 to derive analytical predictions that distinguish the existing accounts and to identify a key empirical signature (a U-shaped allocation of encoding precision) that any successful account must satisfy.

Formally, during the encoding phase, the stimulus p is encoded into a noisy internal signal m, modeled as a one-dimensional transformation with additive Gaussian noise:

$$m = F(p) + \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{5}$$

where $F : [0,1] \to \overline{\mathbb{R}}$ is strictly monotone increasing and smooth. We interpret m as an abstract representation of the neural code for probability. The slope of F at p determines how distinguishable nearby probability values are after encoding: a steeper F stretches a small interval around p into a wider range of m values, making them more separable despite the noise. This is captured by the Fisher Information $\mathcal{J}(p) = (F'(p))^2/\sigma^2$. Following Hahn and Wei [12], we will refer to $\sqrt{\mathcal{J}(p)}$ as the **encoding resources** allocated to p.

During the decoding phase, given m and a prior distribution $P_{\text{prior}}(p)$, Bayesian inference yields a posterior $P(p \mid m)$. We assume the decision-maker derives a point estimate $\hat{p}(m)$ as the posterior mean, which minimizes the Bayes risk for the mean-squared loss function.

In this encoding-decoding framework, the model's behavior is therefore governed by two ingredients: the encoding function F and the prior P_{prior}, both of which we will infer from behavioral data. Of these, the shape of the encoding resources $\sqrt{\mathcal{J}(p)}$ is the central quantity in our analysis, and we will show that the observed inverse-S weighting pattern places strong constraints on its shape. In Section 3.2 we make this correspondence explicit: each existing account is instantiated as a specific choice of F and P_{prior}, and we derive the qualitatively different predictions they make. This sets up the cross-task empirical comparison in Section 4.

3.2 Predictions from the bias decomposition

We analyze the framework's predictions through the bias of the Bayesian estimate, i.e., the average deviation of the estimate $\hat{p}$ from the true probability across trials:

$$\text{Bias}(p) := \mathbb{E}[\hat{p} \mid p] - p \tag{6}$$

Prior work on the bias of Bayesian perceptual models has largely focused on unbounded or circular stimulus spaces, or on stimuli far from a boundary [12, 27, 21, 19, 24]. Probability distortion, by

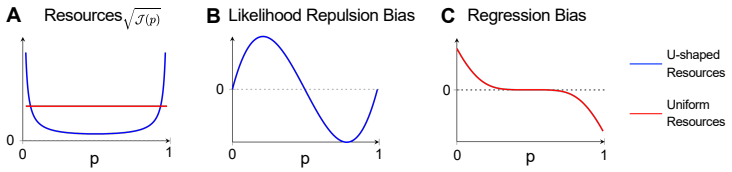


Figure 2: Impact of the encoding (resource allocation) on bias. (A) A U-shaped (blue) and uniform (red) encoding allocation resources $\sqrt{\mathcal{J}(p)}$. (B) The U-shaped encoding generates an S-shaped Likelihood Repulsion bias, $\propto \frac{\mathrm{d}}{\mathrm{d}\,p}(1/\mathcal{J}(p))$. (C) The uniform encoding generates a Regression bias $\propto A_{1,\sigma}(p) \cdot \frac{\mathrm{sign}(0.5-p)}{\sqrt{\mathcal{J}(p)}}$, maximized at 0,1. Both biases generate overestimation of small and underestimation of large probabilities, but with distinct shapes.

contrast, takes place on a bounded interval $[0, 1]$, and the boundaries are not incidental: they are central to longstanding accounts of probability weighting [7, 3].

We extend the Bayesian bias decomposition of Hahn and Wei [12] to the bounded probability interval $[0, 1]$ (Theorem 1; proof in Appendix B.1.1).

Theorem 1. *There are functions $A_{1,\sigma}, A_{2,\sigma}, A_{3,\sigma} : [0,1] \to \mathbb{R}_+$ that are bounded, positive, and satisfy*

$$A_{\ldots,\sigma}(p) \leq \phi\left(\frac{\sigma}{\min\{|F(p)-F(0)|, |F(p)-F(1)|\}}\right) \tag{7}$$

for some nondecreasing function $\phi : \mathbb{R} \to \mathbb{R}$ with $\lim_{x\downarrow 0}\phi(x) = 0$, across all $p \in (0,1); \sigma > 0$. Then, at any $p \in (0,1)$, the Bayesian model has the bias, as $\sigma \to 0$:

$$\mathrm{Bias}(p) = \underbrace{A_{1,\sigma}(p) \cdot \frac{\mathrm{sign}(0.5-p)}{\sqrt{\mathcal{J}(p)}}}_{\textit{Regression from Boundary}} + \underbrace{(1 - A_{2,\sigma}(p)) \cdot \frac{\mathrm{d}}{\mathrm{d}\,p}\left(\frac{1}{\mathcal{J}(p)}\right)}_{\textit{Likelihood Repulsion}}$$
$$+ \underbrace{(1 - A_{3,\sigma}(p)) \cdot \frac{1}{\mathcal{J}(p)} \cdot \frac{\mathrm{d}}{\mathrm{d}\,p}\left(\log P_{prior}(p)\right)}_{\textit{Prior Attraction}} + \mathcal{O}(\sigma^4)$$

The decomposition isolates three distinct mechanisms of distortion: **boundary regression** pushes estimates inward, **likelihood repulsion** pushes them away from regions where encoding resources are concentrated, and **prior attraction** pulls them toward regions of high prior density.

Under this decomposition, the regression-based, log-odds-based, and efficient-coding accounts can be understood within a common structure, each making different assumptions about encoding, priors, and decoding. This common structure allows us to derive predictions that distinguish them empirically, which we develop in the rest of this section and test in Section 4.

The factors $A_{1,\sigma}, A_{2,\sigma}, A_{3,\sigma}$ all vanish when the noise σ is small relative to the distance from $F(p)$ to $F(0)$ and $F(1)$ in the encoding space (Eq. 7). Their role, however, differs across the three terms: far from the boundary, regression vanishes ($A_{1,\sigma} \to 0$), while repulsion and attraction reach full strength ($1 - A_{2,\sigma}, 1 - A_{3,\sigma} \to 1$). Near the boundary, the situation reverses: regression dominates while repulsion and attraction are suppressed.

We examine each in turn, showing how it constrains the encoding F, the prior P_{prior}, or both, and how the resulting predictions can adjudicate among existing accounts.

Boundary regression. The first term in Theorem 1 captures boundary regression. Because probabilities are restricted to $[0, 1]$, the posterior assigns zero probability to values outside this interval. This truncation shifts the posterior mean away from the nearest boundary and toward the center of the interval. Regression-based accounts [7, 3] attribute the inverse-S pattern primarily to boundary regression, treating encoding itself as uniform or otherwise unbiased. Under this view, distortion is a consequence of the bounded response scale rather than of how probabilities are encoded.

This account has a built-in constraint. Boundary regression is strongest near 0 and 1, whereas the inverse-S weighting function vanishes there [22, 25]. The two are compatible only if $\mathcal{J}(0), \mathcal{J}(1) \to$

∞, which is biologically implausible. Accounts that rely primarily on boundary regression therefore predict substantial distortion at the endpoints, in contradiction with the inverse-S pattern. We test this theoretical prediction empirically in Section 4: uniform-encoding accounts should fit the data poorly near the endpoints.

Likelihood repulsion. The second term in Theorem 1 is likelihood repulsion. Where encoding is steep, a small change in the stimulus produces a large change in the internal signal, so only nearby stimuli remain plausible; where it is shallow, even distant stimuli could have produced the same signal. The likelihood extends farther toward the low-precision side, and, when prior effects are weak, so does the posterior, shifting its mean away from regions of high precision.

Consider the first two terms in Theorem 1, boundary regression and likelihood repulsion, independent of the prior. As we showed in the previous paragraph, boundary regression alone cannot produce the inverse-S pattern, which vanishes at $p = 0$ and $p = 1$. When boundary regression is neutralized, the bias is dominated by the likelihood repulsion term $\propto \frac{d}{dp}(1/\mathcal{J}(p))$. This term is S-shaped (Figure 2B) if and only if $\mathcal{J}(p)$ is U-shaped.

Prediction 1. The classic inverse-S weighting pattern implies a U-shaped allocation of encoding resources, with $\sqrt{\mathcal{J}(p)}$ peaking near 0 and 1.

Prediction 1 is consistent with several existing models, although each model arrives at the U-shape resource allocation through a different assumption. Specifically, log-odds-based accounts [30, 17, 18] commit to a fixed log-odds transform $F(p) = \log \frac{p}{1-p}$, which yields a U-shape with $\sqrt{\mathcal{J}(p)} \propto 1/p + 1/(1-p)$. U-shaped-FI accounts [5] allow more general encodings constrained only by U-shaped Fisher Information. Efficient-coding accounts [10] derive the U-shape by matching the encoding to a prior assumed itself to be U-shaped (so that $\sqrt{\mathcal{J}(p)} \propto P_{\text{prior}}(p)$, see Appendix B.3). Within our framework, by contrast, U-shape is the necessary consequence of the inverse-S pattern, requiring no specific functional form on the encoding or prior. In Section 4, we test both points: whether the U-shape can be recovered from data without imposing functional-form constraints, and whether the recovered shapes match the assumptions of these alternative accounts.

As shown in Appendix B.1.2, BLO in the limit of untruncated log-odds is equivalent at the level of bias to a Bayesian model with log-odds encoding and a particular unimodal prior, although it uses a sub-optimal decoder. This connection allows the empirical comparison in Section 4.1 to assess the joint contribution of optimal decoding and flexible prior.

Prior attraction. The third term in Theorem 1 pulls estimates toward regions of high prior density, with strength scaled by $1/\mathcal{J}(p)$. Unlike the previous two terms, prior attraction depends on both P_{prior} and the encoding, so its shape can change as P_{prior} changes—for instance, as the decision-maker is exposed to a new stimulus distribution.

This term is shared, in some form, by all Bayesian accounts [7, 3, 10], but the accounts differ on how flexibly P_{prior} is allowed to vary. BLO [30] approximates prior attraction with a single anchor Λ_0 and does not accommodate general priors. Efficient-coding accounts [10] tie P_{prior} to the encoding through $\sqrt{\mathcal{J}(p)} \propto P_{\text{prior}}(p)$; under this constraint, the posterior mean is biased *away* from the prior mode rather than toward it [26], so prior attraction is effectively absorbed into the encoding. Bayesian accounts with free priors [7, 3] permit attraction toward arbitrary high-density regions.

These assumptions make distinct predictions when the stimulus distribution changes. A flexible Bayesian prior should track the new distribution, producing a corresponding shift in the bias; a single anchor cannot reproduce such shifts; and a prior tied to the encoding can only adapt insofar as the encoding itself adapts, which is generally assumed to occur on slower timescales [8].

Prediction 2. Stimulus distribution changes produce corresponding changes in the bias, mediated by the prior attraction term. Accounts with flexible Bayesian priors predict the largest such effects; accounts with single-anchor priors or prior–encoding matches predict weaker or no adaptation.

We test this prediction in Section 4.3 using a dot-estimation experiment with bimodal stimulus statistics: participants are exposed to a distribution with two modes, and we ask whether their bias adapts in a manner consistent with each account's assumption about the prior.

4 Empirical Validation of the Bayesian Framework

We test the framework's predictions across three behavioral paradigms. In each case, we instantiate the framework with different assumptions about the encoding F and prior P_{prior}, corresponding to the accounts surveyed in Section 2. The judgment of relative frequency (JRF) task tests Prediction 1: whether the inverse-S bias requires non-uniform encoding. Lottery pricing and risky choice test whether the same mechanism extends to economic decisions under risk. Finally, we collected data in a dot-estimation experiment with bimodal stimulus statistics to test Prediction 2: whether bias adapts to changes in the stimulus distribution.

4.1 Validating Prediction 1 on Judgment of Relative Frequency (JRF) data

We analyze data from Zhang and Maloney [29] and Zhang et al. [30], where subjects ($N = 86$) were asked to judge the percentage of dots of a target color in arrays of black and white dots on a gray background.[4]

We compare the parametric baseline **BLO** with three Bayesian variants, each instantiating a different assumption about the encoding while sharing a flexible non-parametric prior. **FreeP+UniformE** assumes uniform encoding $F(p) = p$, instantiating regression-based accounts [7, 3]. **FreeP+BoundedLOE** uses a parametric bounded log-odds encoding, instantiating log-odds-based accounts [17, 18]. **FreeP+FreeE** places no functional form on the encoding, allowing the data to determine its shape. We additionally test the efficient-coding account that prior matches encoding via **FreeP+PriorMatchedE** [10].

Parameters are optimized for each subject individually by maximizing trial-by-trial likelihood; Bayesian models are fit using the method of Hahn and Wei [12], Hahn et al. [13], which represents non-parametric components on a 200-point grid over $[0, 1]$ with smoothness regularization. We primarily compare models using held-out negative log-likelihood (NLL), which evaluates out-of-sample predictive performance irrespective of parameter counts. As a complementary analysis, we use AICc with locally estimated effective degrees of freedom for regularized nonparametric models. Details of both comparison metrics are provided in Appendix C.4.2, and model-fitting procedures are described in Appendix C.4.3.

Model fit shows clear differences across variants (Figure 3D). FreeP+FreeE achieves the lowest summed held-out ΔNLL, outperforming both the parametric baseline BLO and Bayesian variants with constrained encoding (UniformE, BoundedLOE, PriorMatchedE). This advantage is also widespread across participants: FreeP+FreeE achieves lower held-out NLL for 79–83 of the 86 participants in comparisons against UniformE, PriorMatchedE, and BLO (Appendix C.5.3). The advantage persists after accounting for model complexity using summed AICc adjusted by effective degrees offreedom, as shown in Appendix Figure 9. Notably, BoundedLOE, which imposes a parametric U-shaped encoding, improves upon UniformE but performs worse than PriorMatchedE and FreeE in held-out prediction. These results support the importance of nonuniform, U-shaped encoding.

To understand these differences, we compare the model-predicted bias with the observed human bias and examine the bias decomposition (Figure 3B, C). FreeP+UniformE produces a bias driven almost entirely by boundary regression, with zero likelihood repulsion (Panel C, bottom)—by construction, since uniform encoding has constant resources. The resulting fit deviates from the observed human bias near the endpoints (Panel B, bottom). This is consistent with the theoretical constraint discussed in Section 3.2 that uniform encoding cannot reproduce inverse-S weighting. In contrast, FreeP+FreeE better captures the human bias near the boundaries (Panel B, top). Its decomposition is dominated by S-shaped likelihood repulsion, with only minor boundary regression (Panel C, top). Moreover, it recovers a U-shaped encoding with sharp peaks near 0 and 1 (Panel A, top). This pattern emerges without imposing any functional form on F, consistent with Prediction 1.

The previous analyses concern the encoding. The framework's second ingredient, the prior, shows a separate signature in the response variability. Across stimulus values, the variability shows a bimodal pattern with a dip near 0.5 (Appendix Figure 14). Among the variants tested, this structure is captured only by FreeP, whose recovered prior develops a sharp peak around 0.5 because the local high prior density mathematically suppresses variance there. BLO produces an inverted-U variability profile by

[4] A complete description of the tasks and datasets used in this paper is available in Appendix C.2.

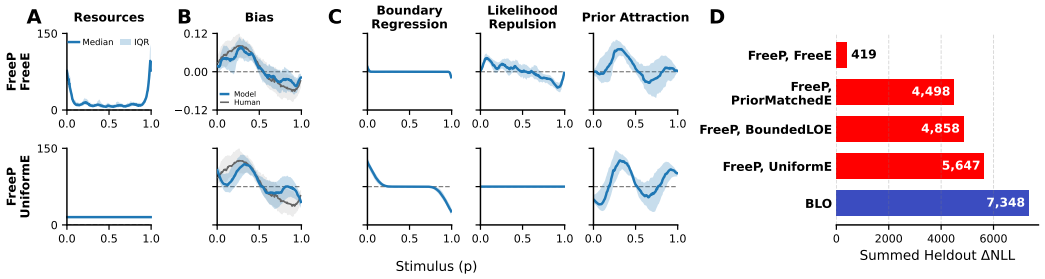


Figure 3: Encoding resources (A), bias (B) and bias decomposition (C) of FreeP+FreeE (top row) and FreeP+UniformE (bottom row) models for JRF data. Lines show medians and shaded regions indicate interquartile ranges (IQR) across subjects. In (B), model-predicted bias (blue) is compared with observed human bias (gray; see Appendix C.11.4 for details). In the best-fitting FreeP FreeE model, boundary regression plays only a small role, whereas Likelihood Repulsion produces a bias away from 0 and 1. The method to compute empirical bias decomposition is described in Appendix C.11.1; results for additional model variants are in Appendix C.9.1. (D) Summed held-out ΔNLL across subjects (lower is better).

construction (Theorem 9 in Appendix B.1.2) and cannot reproduce the bimodal dip. Its single-anchor prior does not allow fine-grained prior structure to shape variability.

The efficient-coding hypothesis predicts that the prior matches the encoding. Enforcing this constraint substantially reduces fit, and the recovered prior and encoding shapes differ in the freely fitted models (Appendix Figure 12 and 13), suggesting that the data require prior and encoding to be specified separately rather than tied together.

Together, these results show that probability distortion in JRF arises from optimal decoding of a U-shaped encoding combined with a flexible, encoding-independent prior. The recovered structure agrees with Prediction 1 even when no functional form is imposed. We next ask whether the same mechanism extends beyond perceptual judgment to economic decision under risk.

4.2 Testing the Generality on Decision-Making Under Risk

Decision-making under risk has been the primary domain for probability weighting research. We test whether the U-shape recovery extends here using two paradigms: lottery pricing and risky choice. We analyze two datasets in which subjects evaluated two-outcome gambles $(x_1, p; x_2, 1-p)$ with real-valued payoffs, instantiated within Cumulative Prospect Theory (CPT) [25]:

$$\text{Utility} := v(x_1) \cdot w(p) + v(x_2) \cdot (1 - w(p)), \quad v(x) = |x|^{\alpha}, \tag{8}$$

with a two-part value function for mixed gambles and separate weighting functions $w^+(p), w^-(p)$ for gains and losses (Appendix Eq. 44). Models differ only in the form of $w(p)$. Parametric accounts (BLO, Prelec, LILO) treat $w(p)$ as a deterministic transformation; in the Bayesian framework, $w(p)$ arises from stochastic encoding and optimal decoding, with FreeP+{UniformE, BoundedLOE, FreeE} instantiating the same family of variants tested in Section 4.1. Both tasks test **Prediction 1**: whether U-shaped encoding resources are recovered in economic decisions, as in perceptual judgment.

Pricing. We first analyze the pricing data from Zhang et al. [30], in which subjects ($N = 75$) chose between a two-outcome monetary gamble and a sure amount, with the sequence of choices adaptively converging to the Certainty Equivalent (CE) of the gamble. Because each CE reflects both the utility function and probability weighting, fitting CE responses directly does not isolate the contribution of $w(p)$. We therefore include an intermediate step: each reported CE is first inverted through the CPT utility function to obtain an implied probability weight, and the Bayesian encoding parameters are estimated from these trial-level weights; the utility exponent α and CE noise variance σ^2_{CE} are then re-optimized to maximize the likelihood of the observed CE data (Appendix C.6.1 for details). Final comparison across all models is based on the likelihood of the original CE responses.

The freely fitted encoding (FreeP+FreeE) recovers a U-shaped allocation of resources (Figure 4 Panel A), confirming Prediction 1 in economic valuation. As in JRF, FreeP+FreeE better captures

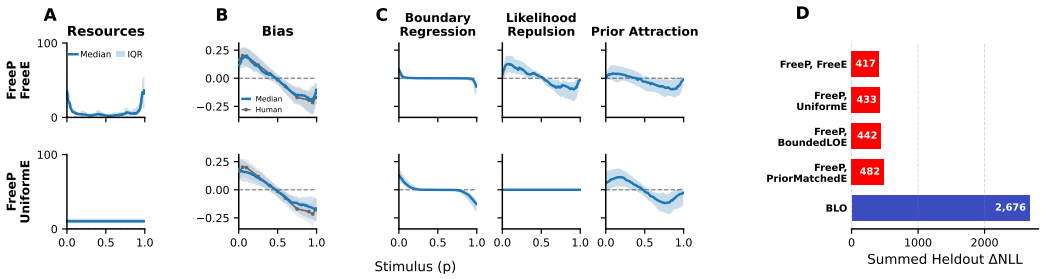


Figure 4: Encoding resources (A), bias (B) and bias decomposition (C) of FreeP+FreeE (top row) and FreeP+UniformE (bottom row) for DMR pricing data. In (B), model-predicted bias (blue) is compared with nonparametrically estimated human bias (gray; see Appendix C.11.4 for details). Lines show medians and shaded regions indicate interquartile ranges (IQR) across subjects. Results for additional model variants are in Appendix C.9.1. (D) Summed held-out ΔNLL across model variants.

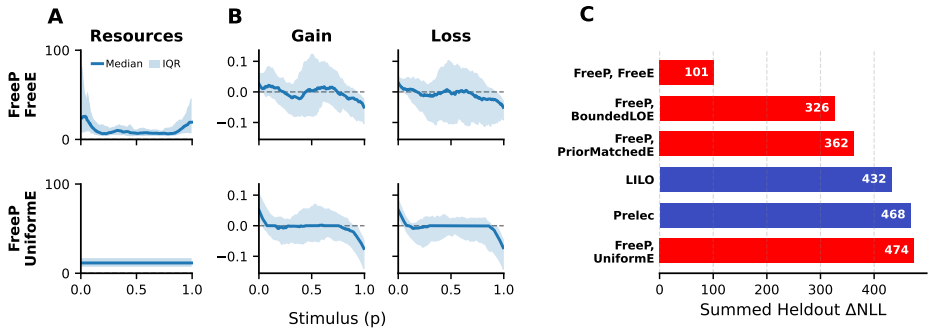


Figure 5: Resources (A), bias (B) and Summed held-out ΔNLL across subjects (C) of FreeP+FreeE and FreeP+UniformE models for DMR Choice data. In the best-fitting FreeP+FreeE model, boundary regression plays only a small role, whereas Likelihood Repulsion produces a bias away from 0 and 1. Results for additional model variants are in Appendix C.9.1.

the observed human bias near the probability boundaries than UniformE (Panel B), with likelihood repulsion playing an important role in shaping the bias (Panel C). FreeP+FreeE achieves the lowest aggregate held-out NLL among the compared models (Panel D). Subject-level comparisons show a clear predictive advantage over BLO, although the differences from UniformE and PriorMatchedE are less conclusive (Appendix C.6.3). We further test the prediction in a choice task below, where responses depend on $w(p)$ directly rather than through the utility function.

Choice. To further test the generality of the framework, we analyzed a subset of the Choice Prediction Competition 2015 dataset [6], restricted to two-outcome gambles with fully described probabilities and uncorrelated payoffs (Amb=0, Corr=0, LotNum=1; 187,150 trials across 153 subjects). Following CPT, probabilities of gains and losses are modeled with separate weighting functions $w^+(p)$ and $w^-(p)$. The Bayesian model implements these by fitting separate priors for gains and losses while sharing the same encoding, with $w^{\pm}(p)$ computed as the expectation of $\hat{p}$ under the encoding distribution. Choice probabilities are modeled with a logit rule on subjective utility differences with a temperature parameter τ. As parametric benchmarks, we include the LILO function (Eq. 45) and the Prelec function [22]; BLO is not applicable here, as its original specification does not extend to choice tasks.

The freely fitted encoding again recovers a U-shaped allocation of resources (Figure 5A). As in JRF and pricing, FreeP+UniformE predicts stronger boundary-regression-driven bias, whereas FreeP+FreeE produces a different bias profile (Panel B). FreeP+FreeE also achieves the best held-out predictive performance among all compared models, including the parametric baselines Prelec and LILO (Panel C). At the subject level, FreeP+FreeE achieves lower held-out NLL for 110–128 of the

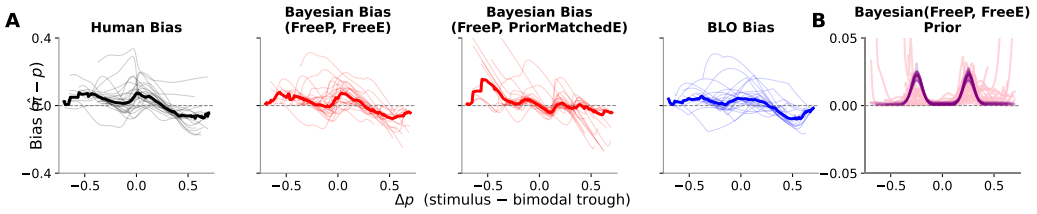


Figure 6: The perceived probability of human observers adapts to bimodal stimulus statistics in a dot-counting task. (A) Median bias across subjects, aligned so the trough between stimulus peaks is at 0 (Δp = stimulus p minus trough). Human bias shows a sign reversal near the higher mode. FreeP+FreeE reproduces this pattern; FreeP+PriorMatchedE predicts the opposite sign of bias near the modes (repulsion rather than attraction); BLO captures the broad shape of the observed bias, with prior attraction represented by a single anchor. (B) Priors recovered by FreeP+FreeE (pink: subjects; purple: ground truth stimulus distribution).

153 participants across the five comparisons (Appendix Table 5). The effective-df AICc comparison also favors FreeP+FreeE (Appendix Figure 26). Within the Bayesian framework, the advantage of FreeE over UniformE provides further support for Prediction 1 in the choice setting.

Across both pricing and choice paradigms, the recovered encoding is U-shaped, agreeing with Prediction 1. We next turn to the second ingredient of the framework, the prior, and test Prediction 2 in an adaptation experiment.

4.3 Testing Prediction 2 via Adaptation to Stimulus Statistics

We now examine the prior component of the framework. Section 3.2 showed that the prior enters bias through a prior-attraction term that pulls estimates toward regions of high prior density. If the prior adapts to the stimulus distribution, a bimodal distribution should induce a bimodal prior, altering the bias through attraction toward the two modes (**Prediction 2**). The accounts reviewed in Section 2 make distinct predictions here. A flexible Bayesian prior can accommodate attraction toward both modes, whereas BLO represents prior attraction through a single anchor. Efficient-coding accounts with matched prior and encoding [10] predict the opposite sign of bias near the modes: when the encoding matches the prior, the posterior mean is biased *away* from the prior mode rather than toward it [26], yielding repulsion rather than attraction.

We tested these predictions in a new psychophysical experiment using the same dot-counting task as in Section 4.1, but with a bimodal stimulus distribution: dot proportions were sampled from a mixture of two Gaussians with peaks equidistant from 0.5. These peaks were chosen separately for each subject.[5] Twenty-six subjects completed the task. Within the Bayesian framework, we extended the set of priors to include a mixture-of-two-Gaussians prior (BimodalP), and added a matched prior-encoding model (FreeP+PriorMatchedE) to instantiate the efficient-coding account.

Human bias under the bimodal distribution shows a local deviation from the inverse-S pattern near the lower mode and a sign reversal near the higher mode (Figure 6A, leftmost). FreeP+FreeE reproduces this pattern (second panel). The prior recovered by FreeP+FreeE is bimodal with peaks aligned to the stimulus modes, closely matching the ground-truth distribution (Figure 6B). This correspondence also holds at the individual-subject level, where recovered priors track the individually assigned stimulus distributions (Appendix Figure 34). This direct recovery of the prior from data agrees with Prediction 2. Notably, the recovered encoding in this experiment remains U-shaped (see Appendix Figure 32), matching the shape recovered in Sections 4.1–4.2 despite the different stimulus distribution. The two components dissociate empirically: the prior adapts while the encoding does not.

BLO captures the broad shape of the observed bias (Figure 6A, rightmost), while its prior-attraction mechanism relies on a single anchor, in contrast to the flexible prior recovered by FreeP+FreeE. FreeP+PriorMatchedE matches the encoding to a bimodal prior, concentrating resources at the modes; through the likelihood-repulsion term, this biases estimates *away* from the modes. Its predicted bias

[5] Full experimental details are in Appendix C.3. The experimental design was approved by the Ethical Review Board of the Faculty of Mathematics and Computer Science, Saarland University.

near each mode (Figure 6A, third panel) has the opposite sign from the human data. Quantitative fit comparisons across all model variants are reported in Appendix Figures 30 and 31. FreeP+FreeE achieves the best aggregate performance under both metrics. At the subject level, FreeP+FreeE achieves lower held-out NLL for all 26 participants compared with BLO and for 18 of 26 participants compared with PriorMatchedE (Appendix Table 6).

These results confirm Prediction 2. Across this section, the encoding component was recovered as U-shaped, and the prior component is recovered as flexible and tracking the stimulus distribution. The two components of the framework are supported by separate empirical tasks. Probability distortion thus emerges from optimal decoding of a U-shaped encoding combined with a flexible Bayesian prior.

5 Discussion and Conclusion

Building on the Bayesian encoding–decoding framework of Hahn and Wei [12], we derive how boundary regression, likelihood repulsion, and prior attraction jointly shape probability estimates on the bounded domain $[0, 1]$. In this framework, the U-shaped allocation of encoding resources is a derived prediction. The bias decomposition also provides a common perspective on existing accounts, highlighting their different assumptions about encoding and prior.

A natural question raised by these results concerns the relationship between prior and encoding. Efficient-coding accounts predict that they should match, but our analyses find them to be empirically distinct in both perceptual and economic tasks. One possibility, consistent with related findings in perception [8], is that priors and encodings adapt on different timescales. The bimodal adaptation experiment is consistent with this view, since the prior shifted within a single session while the recovered encoding remained U-shaped throughout. More broadly, the work connects probability distortion to a growing line of work on imprecise but structured mental representations [e.g. 28, 17, 9, 2, 30, 10]. By showing that a single encoding–decoding mechanism accounts for distortion across both perceptual judgment and economic decision under risk, the framework provides a candidate for what is shared across these domains.

Acknowledgments and Disclosure of Funding

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102 "Information Density and Linguistic Encoding", and Project-ID 560456343. We thank Entang Wang and Yash Sarrof for their helpful discussions and feedback. We also thank the anonymous reviewers for their constructive feedback, which helped improve the paper. The authors declare no competing interests.

References

[1] M. Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: journal of the Econometric Society*, pages 503–546, 1953.

[2] M. Barretto-García, G. de Hollander, M. Grueschow, R. Polania, M. Woodford, and C. C. Ruff. Individual risk attitudes arise from noise in neurocognitive magnitude representations. *Nature Human Behaviour*, 7(9):1551–1567, 2023.

[3] S. Bedi, G. d. Hollander, and C. Ruff. Probability weighting arises from boundary repulsions of cognitive noise. *bioRxiv*, pages 2025–09, 2025.

[4] S. Dehaene. The neural basis of the weber–fechner law: a logarithmic mental number line. *Trends in cognitive sciences*, 7(4):145–147, 2003.

[5] B. Enke and T. Graeber. Cognitive uncertainty. *The Quarterly Journal of Economics*, 138(4): 2021–2067, 2023.

[6] I. Erev, E. Ert, O. Plonsky, D. Cohen, and O. Cohen. From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological review*, 124(4):369, 2017.

[7] J. Fennell and R. Baddeley. Uncertainty plus prior equals rational bias: An intuitive bayesian probability weighting function. *Psychological Review*, 119(4):878, 2012.

[8] M. Fritsche, E. Spaak, and F. P. De Lange. A bayesian and efficient observer model explains concurrent attractive and repulsive history biases in visual perception. *Elife*, 9:e55389, 2020.

[9] C. Frydman and L. J. Jin. Efficient coding and risky choice. *The Quarterly Journal of Economics*, 137(1):161–213, 2022.

[10] C. Frydman and L. J. Jin. On the source and instability of probability weighting. Technical report, National Bureau of Economic Research, 2023.

[11] R. Gonzalez and G. Wu. On the shape of the probability weighting function. *Cognitive psychology*, 38(1):129–166, 1999.

[12] M. Hahn and X.-X. Wei. A unifying theory explains seemingly contradictory biases in perceptual estimation. *Nature Neuroscience*, 27(4):793–804, 2024.

[13] M. Hahn, E. Wang, and X.-X. Wei. Identifiability of bayesian models of perception. *Proceedings of the National Academy of Sciences*, 123(37):e2601013123, 2026. doi: 10.1073/pnas.2601013123. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2601013123`.

[14] J. A. Heng, M. Woodford, and R. Polania. Efficient sampling and noisy decisions. *Elife*, 9: e54962, 2020.

[15] K. Juechems, J. Balaguer, B. Spitzer, and C. Summerfield. Optimal utility and probability functions for agents with finite computational precision. *Proceedings of the National Academy of Sciences*, 118(2):e2002232118, 2021.

[16] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):363–391, 1979.

[17] M. W. Khaw, Z. Li, and M. Woodford. Cognitive imprecision and small-stakes risk aversion. *The review of economic studies*, 88(4):1979–2013, 2021.

[18] M. W. Khaw, Z. Li, and M. Woodford. Cognitive imprecision and stake-dependent risk attitudes. CESifo Working Paper Series 9923, CESifo, 2022. URL `https://ideas.repec.org/p/ces/ceswps/_9923.html`.

[19] M. Morais and J. W. Pillow. Power-law efficient neural codes provide general link between perceptual bias and discriminability. *Advances in Neural Information Processing Systems*, 31, 2018.

[20] F. H. Petzschner and S. Glasauer. Iterative bayesian estimation as an explanation for range and regression effects: a study on human path integration. *Journal of Neuroscience*, 31(47): 17220–17229, 2011.

[21] A. Prat-Carrabin and M. Woodford. Bias and variance of the bayesian-mean decoder. *Advances in neural information processing systems*, 34:23793–23805, 2021.

[22] D. Prelec. The probability weighting function. *Econometrica*, pages 497–527, 1998.

[23] K. Ruggeri, S. Alí, M. L. Berge, G. Bertoldo, L. D. Bjørndal, A. Cortijos-Bernabeu, C. Davison, E. Demić, C. Esteban-Serna, M. Friedemann, et al. Replicating patterns of prospect theory for decision under risk. *Nature human behaviour*, 4(6):622–633, 2020.

[24] A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, 9(4):578–585, 2006.

[25] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

[26] X.-X. Wei and A. A. Stocker. A bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts. *Nature neuroscience*, 18(10):1509–1517, 2015.

[27] X.-X. Wei and A. A. Stocker. Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, 114(38):10244–10249, 2017.

[28] M. Woodford. Prospect theory as efficient perceptual distortion. *American Economic Review*, 102(3):41–46, 2012.

[29] H. Zhang and L. T. Maloney. Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, Volume 6 - 2012, 2012. ISSN 1662-453X. doi: 10.3389/fnins.2012.00001. URL `https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2012.00001`.

[30] H. Zhang, X. Ren, and L. T. Maloney. The bounded rationality of probability distortion. *Proceedings of the National Academy of Sciences*, 117(36):22024–22034, 2020. doi: 10.1073/pnas.1922401117. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1922401117`.

Appendix

A FAQs **14**

B Theoretical Derivations and Proofs **15**

B.1 Theory on Fisher Information (FI), Bias and Mean Square Error 15

B.1.1 Properties of Bayesian Model 16

B.1.2 Properties of BLO Model 19

B.2 Bayesian Models with Log-Odds Encoding 21

B.3 FI for Efficient Code in Frydman and Jin [10] 22

C Experimental Methods and Supplementary Results **22**

C.1 Theoretical Mapping of Model Variants 22

C.2 Task and Datasets 23

C.2.1 Judgment of Relative Frequency (JRF) 23

C.2.2 Pricing Task in Decision-making Under Risk (DMR) 23

C.2.3 Choice Task in Decision-making Under Risk (DMR) 23

C.3 Adaptation Experiment 24

C.3.1 Recruitment and Task instructions 24

C.4 Model Fitting and Evaluation 25

C.4.1 Implementation and Optimization Details 25

C.4.2 Model Comparison Metrics 25

C.4.3 Categorical and Analytical Fitting 26

C.5 Details on Bayesian Model for JRF Task 28

C.5.1 Bayesian Model Variants 28

C.5.2 Performance Comparison of All Model Variants 28

C.5.3 Subject-level Held-out Model Comparisons 30

C.5.4 Analysis of Fitted Prior and Resources 30

C.5.5 Analysis of Bias and Variance 32

C.5.6 Sensitivity to Grid Size, Regularization Strength, and Likelihood Specification 34

C.5.7 Sensitivity to the Decoding Rule 34

C.5.8 Encoding Recovery Analysis 36

C.6 Details on Bayesian Model for DMR Pricing Task 36

C.6.1 Bayesian Model Variants and Fitting Procdure 36

C.6.2 Performance Comparison of All Model Variants 38

C.6.3 Subject-level Held-out Model Comparisons 38

C.6.4 Analysis of Fitted Prior and Resources 39

C.6.5 Analysis of Bias and Variance 40

C.7 Details on Bayesian Model for DMR Choice Task 40

C.7.1 Full Utility Function, Logit Choice Rule, and LILO function 40

C.7.2 Bayesian Model Variants and Fitting Procedure . . . 41
C.7.3 Performance Comparison of Model Variants . . . 41
C.7.4 Subject-level Held-out Model Comparisons . . . 41
C.7.5 Analysis of Fitted Prior and Resources . . . 42
C.7.6 Analysis of Bias . . . 42
C.8 Details on Bayesian Model for Adaptation Bimodal Data . . . 43
C.8.1 Bayesian Model Variants . . . 43
C.8.2 Performance Comparison of All Model Variants . . . 43
C.8.3 Subject-level Held-out Model Comparisons . . . 45
C.8.4 Robustness of Model Comparison via Bootstrapping . . . 45
C.8.5 Analysis of Fitted Prior and Resources . . . 46
C.8.6 Analysis of Bias and Variance . . . 47
C.8.7 Sensitivity to the Decoding Rule . . . 48
C.9 Cross-Task Analysis . . . 49
C.9.1 Bias Decomposition . . . 49
C.9.2 Resources and Priors . . . 49
C.10 Implementation Details of BLO Baselines . . . 49
C.10.1 JRF and Adapation . . . 49
C.10.2 DMR Pricing . . . 52
C.11 Additional Analysis and Plotting Methods . . . 53
C.11.1 Methods for Generating Empirical Bias Decomposition . . . 53
C.11.2 Validity of the Small-Noise Approximation . . . 54
C.11.3 Details for Plotting Group-level Prior and Resources . . . 54
C.11.4 Methods for Calculating Bias and Variance . . . 54

A FAQs

1. *What is the contribution beyond Hahn and Wei [12], Hahn et al. [13] or existing Bayesian accounts?*

   Hahn and Wei [12] and Hahn et al. [13] provide general-purpose tools for Bayesian encoding–decoding models on unbounded or circular domains. We build on this infrastructure to address a specific open question in behavioral science: where does the probability weighting function come from?

   Our contribution has three parts: First, probability lives on [0,1] with two hard boundaries. In Theorem 1, we derive a bias decomposition that accounts for both boundaries simultaneously, with boundary-modulated coefficients describing how regression, repulsion, and attraction are affected near each endpoint. Second, we establish that the inverse-S weighting pattern arises from a U-shaped allocation of encoding precision, recovered nonparametrically across perceptual estimation, lottery pricing, and risky choice, and stable under changes in stimulus distribution. Third, we show that regression-based, log-odds, efficient-coding, and BLO accounts emerge as special cases of a single framework, and the bias decomposition yields testable predictions that distinguish them.

2. *FreeP+FreeE is more flexible than other models. Isn't it unsurprising that it outperforms other models?*

   First, we emphasize that the comparison uses two metrics that control for capacity: Held-out negative log-likelihood penalizes overfitting directly; AICc with effective degree of freedom

penalizes flexibility (see Appendix Section C.4.2). Importantly, FreeP+FreeE wins on both despite its increased flexibility. Second, notably, across perceptual estimation, lottery pricing, and risky choice, free fitting consistently recovers a U-shape, and the same shape persists when the stimulus distribution becomes bimodal. The convergence on a specific encoding geometry is a key prediction of the framework.

3. *The adaptation experiment (N=26) is substantially smaller than the JRF (N=86) and pricing (N=75) datasets. Are the conclusions robust at this sample size?*

   Although the adaptation sample is smaller than those of the other tasks, participant-level bootstrap analyses provide additional support for the aggregate predictive results. Across 10,000 bootstrap resamples, FreeP+FreeE achieved lower summed held-out NLL than PriorMatchedE in 98.1% of samples and than BLO in 100%. See Appendix C.8.4 for full results.

4. *Existing log-odds and efficient-coding accounts already assume a U-shaped FI; what is new here?*

   These accounts posit a U-shape as a theoretical assumption: log-odds accounts derive it from a fixed $\log \frac{p}{1-p}$ transform, and efficient-coding accounts derive it from matching the encoding to a U-shaped prior. In contrast, we recover the U-shape nonparametrically and without any functional-form assumption on F. By formulating the various existing accounts as special cases of a common framework, we are able to ask which modeling assumptions are supported by the data. First, pure log-odds encoding implies $\sqrt{\mathcal{J}(p)} \to \infty$ at the boundaries, which is biologically implausible; BLO addresses this by truncating log-odds, but the truncation introduces discontinuities in the encoding resources. In contrast, the recovered shape has finite precision at the end points and remains smooth throughout. Second, efficient coding ties the U-shape to the prior, whereas the recovered encoding remains U-shaped even when the prior becomes bimodal under adaptation. Taken together, our analyses show that the data favor a U-shaped encoding combined with a free prior over any of the constrained alternatives.

5. *Why would the brain use a U-shaped encoding?*

   In Appendix B.2, we describe a possible motivation of U-shaped encodings grounded in the well-documented encoding of scalar quantities following Weber's law.

B Theoretical Derivations and Proofs

B.1 Theory on Fisher Information (FI), Bias and Mean Square Error

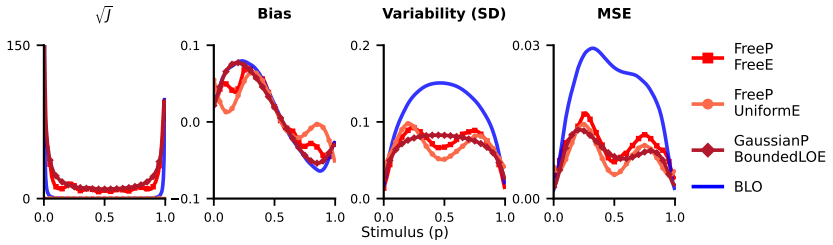


Figure 7: FI, Bias, Variance, MSE of the three models: Bayesian, BLO, Bayesian with uniform encoding, for parameters fitted in Section 4.1. BLO has an FI peaking at 0 and 1, as does the Bayesian model with log-odds encoding. All model variants predict a positive bias for small p, and a negative bias for large p, in line with the classical probability weighting function (Figure 1A). BLO and models with Gaussian priors predict a simple inverse U-shaped variability; other Bayesian priors can accommodate other variabilities.

B.1.1 Properties of Bayesian Model

Theorem 2 (Repeated from Theorem 1)**.** *There are functions $A_{1,\sigma}, A_{2,\sigma}, A_{3,\sigma} : [0,1] \to \mathbb{R}$ that are bounded, positive, and satisfy*

$$A_{\dots,\sigma}(p) \leq \phi\left(\frac{\sigma}{\min\{|F(p) - F(0)|, |F(p) - F(1)|\}}\right) \tag{9}$$

for some nondecreasing function $\phi : \mathbb{R} \to \mathbb{R}$ with $\lim_{x \downarrow 0} \phi(x) = 0$, across $p \in (0,1); \sigma > 0$. Then, at any $p \in (0,1)$, the Bayesian model has the bias, as $\sigma \to 0$:

$$\begin{aligned}\mathrm{Bias}(p) = &\underbrace{A_{1,\sigma}(p) \cdot \frac{\mathrm{sign}(0.5 - p)}{\sqrt{\mathcal{J}(p)}}}_{\textit{Regression from Boundary}} + \underbrace{(1 - A_{2,\sigma}(p)) \cdot \frac{\mathrm{d}}{\mathrm{d}\,p}\left(\frac{1}{\mathcal{J}(p)}\right)}_{\textit{Likelihood Repulsion}} \\ &+ \underbrace{(1 - A_{3,\sigma}(p))\frac{1}{\mathcal{J}(p)} \cdot \frac{\mathrm{d}}{\mathrm{d}\,p}\left(\log P_{prior}(p)\right)}_{\textit{Prior Attraction}} + \mathcal{O}(\sigma^4)\end{aligned}$$

Proof. We first consider the case where $F(0) = -\infty$, $F(1) = \infty$, in which case (9) forces $A_{\dots,\sigma}(p) \equiv 0$. The above expression with $A_{\dots,\sigma}(p) \equiv 0$ simplifies to:

$$\underbrace{\frac{\mathrm{d}}{\mathrm{d}\,p}\left(\frac{1}{\mathcal{J}(p)}\right)}_{\text{Likelihood Repulsion}} + \underbrace{\frac{1}{\mathcal{J}(p)} \cdot \frac{\mathrm{d}}{\mathrm{d}\,p}\left(\log P_{prior}(p)\right)}_{\text{Prior Attraction}} + \mathcal{O}(\sigma^4)$$

Indeed, this is exactly the expression from Theorem 1 in Hahn and Wei [12]; the claim thus follows from that theorem.

We next consider the case where $F(0), F(1)$ are finite; we obtain this on the basis of Theorem 3 in Hahn and Wei [12]. That theorem examines the bias in the setting where the stimulus space has *one* boundary (e.g., $(-\infty, \theta_{Max}]$). For this case, it proves the decomposition (we consider the special case of $p = 2$, as we assume the posterior mean estimator):

$$\underbrace{-\frac{C_{1,p,D,F,\sigma}}{\sqrt{\mathcal{J}(p)}}}_{\text{Regression from Boundary}} + \underbrace{C_{3,p,D}\left(\frac{1}{\mathcal{J}(p)}\right)'}_{\text{Likelihood Repulsion}} + \underbrace{C_{2,p,D}\frac{1}{\mathcal{J}(p)}\left(\log P_{prior}(p)\right)'}_{\text{Prior Attraction}} + \mathcal{O}(\sigma^4) \tag{10}$$

where $D(p) := \frac{F(\theta_{Max}) - F(p)}{\sigma}$, and where $C_{\dots}$ are given as follows:

1. $C_{1,p,D,F,\sigma}$ depends on the probability p, the quantity $D(p)$, the encoding function F, and the sensory noise magnitude σ
2. $C_{2,p,D}$ depends on the probability p and the quantity $D(p)$
3. $C_{3,p,D}$ depends on the probability p and the quantity $D(p)$
4. $C_{1,p,D,F,\sigma}$, $C_{2,p,D}$, $C_{3,p,D}$ are positive everywhere
5. $C_{1,p,D,F,\sigma} = \Theta(1)$ when $\sigma \to 0$
6. $C_{1,\dots} \to_{D \to \infty} 0$, $C_{2/3,\dots} \to_{D \to \infty} 1$.

In fact, inspecting the proof of the theorem shows that $C_{3,p,D}, C_{2,p,D} \in (0,1)$ for all values of p, D, and that $C_{1,p,D(p),F,\sigma}$ is continuous in p and σ for $\sigma > 0$.

Our aim is to use this result to establish the bias in the setting where the stimulus space is $[0,1]$, i.e., has *two* boundaries.

Towards this end, we partition $(0,1)$ into $(0,1/2)$ and $(1/2,1)$.

First, $p \in (1/2,1)$, the above decomposition is valid with $\theta_{Max} = 1$. The reason is because the contribution of the lower boundary at 0 to the bias can be absorbed into the $\mathcal{O}(\sigma^4)$ remainder, because

only exponentially small probability mass of the likelihood of the encoding m falls into regions outside a small local environment of p [12, SI Appendix, Section S.3.1.1].

Second, for $p \in (0, 1/2)$, by transforming the above result via $p \mapsto -p$, we obtain for the stimulus space $[\theta_{Min}, \infty)$ the bias:

$$\underbrace{\frac{C_{1,p,D,F,\sigma}}{\sqrt{\mathcal{J}(p)}}}_{\text{Regression from Boundary}} + \underbrace{C_{3,p,D}\left(\frac{1}{\mathcal{J}(p)}\right)'}_{\text{Likelihood Repulsion}} + \underbrace{C_{2,p,D}\frac{1}{\mathcal{J}(p)}\left(\log P_{prior}(p)\right)'}_{\text{Prior Attraction}} + \mathcal{O}(\sigma^4) \tag{11}$$

where $D(p) := \frac{F(p)-F(\theta_{Min})}{\sigma}$, and where $C_{...}$ satisfy exactly the same conditions as given above.

Now write $\hat{C}_{...}$ for the coefficients for $p \in (1/2, 1)$ and $\tilde{C}_{...}$ for the coefficients for $p \in (0, 1/2)$; the same with $\hat{D}(p)$ and $\tilde{D}(p)$.

Now define

$$\begin{aligned}
A_{1,\sigma}(p) &:= \hat{C}_{1,p,D(p),F,\sigma} \cdot 1_{p \geq 1/2} + \tilde{C}_{1,p,D(p),F,\sigma} \cdot 1_{p<1/2} \\
A_{2,\sigma}(p) &:= 1 - \left(\hat{C}_{3,p,D(p)} \cdot 1_{p \geq 1/2} + \tilde{C}_{3,p,D(p)} \cdot 1_{p<1/2}\right) \\
A_{3,\sigma}(p) &:= 1 - \left(\hat{C}_{2,p,D(p)} \cdot 1_{p \geq 1/2} + \tilde{C}_{2,p,D(p)} \cdot 1_{p<1/2}\right)
\end{aligned}$$

First, inserting these definitions shows that the two decompositions do indeed produce the one claimed by the theorem.

We now need to show the remaining conclusions. By the assumptions above, $A_{1,\sigma}$ is positive and also continuous in both σ and p. Next, by the facts that $C_{1,p,D,F,\sigma} = \Theta(1)$ when $\sigma \to 0$ and $C_{1,...} \to_{D\to\infty} 0$, we find that $A_{1,\sigma}(p)$ is bounded across σ and p whenever σ is upper-bounded by some σ_{Max}. As $C_{2,...}, C_{3,...} \in [0, 1]$, we find that $A_{2,\sigma}$ and $A_{3,\sigma}$ are positive and bounded. We have thus shown that $A_{1,\sigma}$, $A_{2,\sigma}$, $A_{3,\sigma}$ are positive and bounded.

Finally, as mentioned above, $\hat{C}_{1,p,\hat{D}(p),F,\sigma} \to 0$ as $\hat{D}(p) \to \infty$; equivalently,

$$\hat{C}_{1,p,\hat{D}(p),F,\sigma} = \mathcal{O}\left(\phi_1\left(\frac{1}{\hat{D}(p)}\right)\right) \tag{12}$$

for some nondecreasing function ϕ_1 with $\lim_{s\downarrow 0} \phi_1(s) = 0$. The same holds for $\tilde{C}$, $\tilde{D}$ with a function ϕ_2. Thus,

$$\begin{aligned}
A_{1,\sigma}(p) =& \mathcal{O}\left(\phi_1\left(\frac{1}{\hat{D}(p)}\right)\right) + \mathcal{O}\left(\phi_2\left(\frac{1}{\tilde{D}(p)}\right)\right) \\
=& \mathcal{O}\left(\max\left\{\phi_1(\frac{1}{\hat{D}(p)}), \phi_2(\frac{1}{\tilde{D}(p)})\right\}\right)
\end{aligned}$$

which is equivalent to the claim that

$$A_{1,\sigma}(p) \leq \phi\left(\frac{\sigma}{\min\{|F(p) - F(0)|, |F(p) - F(1)|\}}\right) \tag{13}$$

across $p \in (0, 1); \sigma > 0$, for some nondecreasing function ϕ with $\lim_{s\downarrow 0} \phi(s) = 0$. An analogous argument applies to $A_{2,\sigma}$ and $A_{3,\sigma}$.

□

Extension to Posterior L_q Loss. Theorem 1 extends to posterior L_q-risk minimizers with even integers $q \geq 2$, defined as

$$\hat{p}_q(m) = \arg\min_{a\in[0,1]} \mathbb{E}_{X\sim P(X|m)}\left[(a - X)^q\right]. \tag{14}$$

The bias decomposition retains the same three components, with the likelihood-repulsion term multiplied by $(q+2)/4$:

$$\begin{aligned}\mathrm{Bias}_q(p) = & \underbrace{A_{1,\sigma}^{(q)}(p)\frac{\mathrm{sign}(0.5-p)}{\sqrt{\mathcal{J}(p)}}}_{\text{Boundary Regression}} \\ & + \underbrace{\left(1-A_{2,\sigma}^{(q)}(p)\right)\frac{q+2}{4}\frac{\mathrm{d}}{\mathrm{d}p}\left(\frac{1}{\mathcal{J}(p)}\right)}_{\text{Likelihood Repulsion}} \\ & + \underbrace{\left(1-A_{3,\sigma}^{(q)}(p)\right)\frac{1}{\mathcal{J}(p)}\frac{\mathrm{d}}{\mathrm{d}p}\log P_{\mathrm{prior}}(p)}_{\text{Prior Attraction}} + \mathcal{O}(\sigma^4).\end{aligned}$$

The extension follows from the general L_q results of Hahn and Wei [12]. The same two-boundary argument used in the proof of Theorem 1 applies, with the likelihood-repulsion coefficient changing to $(q+2)/4$ and the boundary correction functions depending on q.

Since $(q+2)/4 > 0$, the likelihood-repulsion term retains its sign and dependence on the encoding resources. Thus, the prediction that U-shaped encoding resources produce S-shaped likelihood repulsion is not specific to posterior-mean decoding.

Corollary 3. *For unbounded log-odds encoding:*

$$\mathbb{E}[\hat{p}|p] - p = \underbrace{\sigma^2 p^2(1-p)^2\left(\log P_{prior}(p)\right)'}_{\textit{Attraction}} + \underbrace{2\sigma^2 p(1-p)(1-2p)}_{\textit{Likelihood Repulsion}} + \mathcal{O}(\sigma^4) \tag{15}$$

Proof. When $\beta = 0$, (34) simplifies to:

$$\sqrt{\mathcal{J}(p)} = \frac{1}{\sigma \cdot p \cdot (1-p)} \tag{16}$$

Furthermore, the coefficients $A_{\dots}$ vanish for this encoding, because the encoding map $F(p) = \int_{1/2}^{p}\sqrt{\mathcal{J}(q)}dq$ is the log-odds transformation, satisfying $F(0) = -\infty$, $F(1) = \infty$. Now

$$\frac{1}{\mathcal{J}(p)} = \sigma^2 \cdot p^2 \cdot (1-p)^2 \tag{17}$$

with derivative

$$\left(\frac{1}{\mathcal{J}(p)}\right)' = \sigma^2 \cdot 2 \cdot p \cdot (1-p) \cdot (1-2p) \tag{18}$$

Plugging these into Theorem 1 yields the result.

□

Theorem 4. *At each $p \in (0,1)$, the Bayesian model has the response variability:*

$$\mathrm{Var}[\hat{p} \mid p] = \frac{1}{\mathcal{J}(p)} + O(\sigma^4). \tag{19}$$

Proof. See Lemma S24 in Hahn et al. [13]. □

Theorem 5 (Optimality of Decoding)**.** *The MSE of the decoded estimate in the Bayesian Model is*

$$\frac{1}{\mathcal{J}(p)} + O(\sigma^4) \tag{20}$$

Proof. Immediate, from $MSE = Var + Bias^2$, and noting that $Bias = \mathcal{O}(\sigma^2)$. □

B.1.2 Properties of BLO Model

Theorem 6 (Repeated from Theorem 10)**.** *The BLO model in the limit of untruncated log-odds, and a Bayesian model with unbounded log-odds encoding ($\beta = 0$) and a specific unimodal prior $P_{prior}(p)$ (depending on Λ_0) have the same bias up to difference $\mathcal{O}(\kappa^2 + \sigma^4)$.*

Proof. We derive this as a corollary of Theorem 8: At any $p \in (0, 1)$, the BLO model has the bias:

$$\Phi(p) - p + \underbrace{\kappa \cdot \Phi(p)(1 - \Phi(p))V_p[\Lambda_0 - \Lambda(p)]}_{\equiv \text{Attraction}} + \underbrace{\frac{\sigma^2}{2}\Phi(p)\left(1 - \Phi(p)\right)\left(1 - 2\Phi(p)\right)}_{\equiv \text{Repulsion}} + O(\kappa^2) + O(\sigma^4) \tag{21}$$

where $\Phi : (0, 1) \to (0, 1) : \Phi(p) := \lambda^{-1}(\hat{\Lambda}(p))$. Now, in the limit of untruncated log-odds:

$$\begin{aligned} &\Delta_+ \to \infty \\ &\Delta_- \to -\infty \\ &\Psi \to \infty \text{ while } \frac{\Psi}{\Delta_+ - \Delta_-} \equiv \text{const} \end{aligned}$$

the value $\Phi(p)$ equals p. We now obtain the result by matching (35) and (26). Specifically, (26) then assumes the form:

$$\underbrace{\kappa \cdot p(1 - p)V_p[\Lambda_0 - \Lambda(p)]}_{\equiv Attraction} + \underbrace{\frac{\sigma^2}{2}p\left(1 - p\right)\left(1 - 2p\right)}_{\equiv Repulsion} + O(\kappa^2) + O(\sigma^4) \tag{22}$$

We match this with the Bayesian bias:

$$\underbrace{\sigma^2 p^2(1 - p)^2 \left(\log P_{prior}(p)\right)'}_{\text{Attraction}} + \underbrace{2\sigma^2 p(1 - p)(1 - 2p)}_{\text{Likelihood Repulsion}} + \mathcal{O}(\sigma^4) \tag{23}$$

Under the identification $\kappa = \sigma^2_{Bayesian}$, $\sigma^2_{BLO} = 4\sigma^2_{Bayesian}$, we find

$$\left(\log P_{prior}(p)\right)' = \frac{V_p[\Lambda_0 - \Lambda(p)]}{p(1 - p)} \propto \Lambda_0 - \Lambda(p) \tag{24}$$

Since $\Lambda(p)$ is monotonically increasing in p, this quantity is monotonically decreasing. Hence, $P_{prior}(p)$ has a single peak and describes a unimodal prior. □

Theorem 7. *The BLO model has resource allocation $\sqrt{\mathcal{J}(p)}$ equal to:*

$$\frac{1}{\sigma} \cdot \left[\left(-\kappa \frac{1 - 2p}{(1 + \kappa p(1 - p))^2}\right) \cdot (\Lambda(p) - \Lambda_0) + \frac{\omega_p \cdot \Psi}{(\Delta_+ - \Delta_-)/2} \begin{cases} 0 & if \ \ \lambda(p) \notin [\Delta_-, \Delta_+] \\ \lambda\prime(p) & else \end{cases}\right] \tag{25}$$

Proof of Theorem 7. Given the standard expression for the Fisher information of a Gaussian with parameter-dependent mean and constant variance, $\sqrt{\mathcal{J}(p)}$ equals

$$\begin{aligned}
&\frac{1}{\sigma}\cdot\frac{d\hat{\Lambda}_\omega(p)}{dp}\\
=&\frac{1}{\sigma}\cdot\frac{d}{dp}\left[\omega_p\cdot\Lambda(p)+(1-\omega_p)\cdot\Lambda_0\right]\\
=&\frac{1}{\sigma}\cdot\frac{d}{dp}\left[\omega_p\cdot\Lambda(p)-\omega_p\cdot\Lambda_0\right]\\
=&\frac{1}{\sigma}\cdot\left[\left(\frac{d}{dp}\omega_p\right)\cdot\Lambda(p)+\omega_p\left(\cdot\frac{d}{dp}\Lambda(p)\right)-\Lambda_0\cdot\left(\frac{d}{dp}\omega_p\right)\right]\\
=&\frac{1}{\sigma}\cdot\left[\left(\frac{d}{dp}\omega_p\right)\cdot(\Lambda(p)-\Lambda_0)+\omega_p\left(\cdot\frac{d}{dp}\Lambda(p)\right)\right]\\
=&\frac{1}{\sigma}\cdot\left[\left(\frac{d}{dp}\frac{1}{1+\kappa p(1-p)}\right)\cdot(\Lambda(p)-\Lambda_0)+\omega_p\cdot\left(\frac{\Psi}{(\Delta_+-\Delta_-)/2}\cdot\frac{d}{dp}\Gamma(\lambda(p))\right)\right]\\
=&\frac{1}{\sigma}\cdot\left[\left(-\kappa\frac{1-2p}{(1+\kappa p(1-p))^2}\right)\cdot(\Lambda(p)-\Lambda_0)+\frac{\omega_p\cdot\Psi}{(\Delta_+-\Delta_-)/2}\begin{cases}0 & if\ \lambda(p)\notin[\Delta_-,\Delta_+]\\ \lambda\prime(p) & else\end{cases}\right]
\end{aligned}$$

□

Theorem 8. *At any $p\in(0,1)$, the BLO model has the bias:*

$$\Phi(p)-p+\underbrace{\kappa\cdot\Phi(p)(1-\Phi(p))V_p[\Lambda_0-\Lambda(p)]}_{\equiv Attraction}+\underbrace{\frac{\sigma^2}{2}\Phi(p)\left(1-\Phi(p)\right)\left(1-2\Phi(p)\right)}_{\equiv Repulsion}+O(\kappa^2)+O(\sigma^4) \tag{26}$$

where $\Phi:(0,1)\to(0,1):\Phi(p):=\lambda^{-1}(\hat{\Lambda}(p))$.

Proof of Theorem 8. Consider the estimate as a function of p:

$$\hat{\pi}(p)=\lambda^{-1}(\hat{\Lambda}_\omega(p)+\epsilon_\lambda) \tag{27}$$

The bias is its expectation over ϵ_λ minus the true value:

$$\mathbb{E}\left[\lambda^{-1}(\hat{\Lambda}_\omega(p)+\epsilon_\lambda)\right]-p \tag{28}$$

where $\epsilon_\lambda\sim N(0,\sigma^2)$. To understand it, we perform a Taylor expansion around $\sigma=0$, $\kappa=0$. That is, we start by computing

$$\frac{\partial^2}{(\partial\epsilon_\lambda)^2}\left[\lambda^{-1}(\hat{\Lambda}_\omega(p)+\epsilon_\lambda)\right]=\lambda^{-1}(\Lambda(p))\left(1-\lambda^{-1}(\Lambda(p))\right)\left(1-2\lambda^{-1}(\Lambda(p))\right) \tag{29}$$

at $\epsilon_\lambda=0$, $\kappa=0$, and

$$\frac{\partial}{\partial\kappa}\left[\lambda^{-1}(\hat{\Lambda}_\omega(p)+\epsilon_\lambda)\right]=\lambda^{-1}(\Lambda(p))\left(1-\lambda^{-1}(\Lambda(p))\right)V_p\left[\Lambda_0-\Lambda(p)\right] \tag{30}$$

at $\kappa=0$, $\epsilon_\lambda=0$.

Then

$$\begin{aligned}
\mathbb{E}\left[\lambda^{-1}(\hat{\Lambda}_\omega(p)+\epsilon_\lambda)\right]-p=&\kappa\cdot\frac{\partial}{\partial\kappa}\left[\lambda^{-1}(\hat{\Lambda}_\omega(p)+\epsilon_\lambda)\right]\\
&+\frac{\sigma^2}{2}\frac{\partial^2}{(\partial\epsilon_\lambda)^2}\left[\lambda^{-1}(\hat{\Lambda}_\omega(p)+\epsilon_\lambda)\right]\\
&+\lambda^{-1}(\Lambda(p))\\
&-p\\
&+O(\kappa^2)+O(\sigma^4)
\end{aligned}$$

Filling in the above expressions, we get

$$\begin{aligned}\mathbb{E}\left[\lambda^{-1}(\hat{\Lambda}_\omega(p)+\epsilon_\lambda)\right]-p =&\kappa\cdot\lambda^{-1}(\Lambda(p))(1-\lambda^{-1}(\Lambda(p)))V_p[\Lambda_0-\Lambda(p)]\\ &+\frac{\sigma^2}{2}\lambda^{-1}(\Lambda(p))\left(1-\lambda^{-1}(\Lambda(p))\right)\left(1-2\lambda^{-1}(\Lambda(p))\right)\\ &+\lambda^{-1}(\Lambda(p))\\ &-p\\ &+O(\kappa^2)+O(\sigma^4)\end{aligned}$$

□

Theorem 9. *Given fixed p, let $q := \lambda^{-1}(\hat{\Lambda}_\omega(p))$. Then the BLO model has the response variance:*

$$\mathrm{Var}[\hat{\pi}(p)\mid p] = q^2(1-q)^2\,\sigma^2 + q^2(1-q)^2\left(\frac{3}{2}-8q+8q^2\right)\sigma^4+O(\sigma^6).$$

Proof. Consider the estimate as a function of p:

$$\hat{\pi}(p)=\lambda^{-1}(\hat{\Lambda}_\omega(p)+\epsilon_\lambda) \tag{31}$$

Conditioning on p, the variance over ϵ_λ is using a Taylor expansion:

$$\mathrm{Var}\left[\lambda^{-1}(\hat{\Lambda}_\omega(p)+\epsilon_\lambda)\right]=\left(f'(\mu)\right)^2\sigma^2+f'(\mu)f'''(\mu)\sigma^4+\tfrac{1}{2}\left(f''(\mu)\right)^2\sigma^4+O(\sigma^6).$$

where $\mu := \hat{\Lambda}_\omega(p)$ and $f(x) := \lambda^{-1}(x)$. Plugging this in, we obtain the result. □

B.2 Bayesian Models with Log-Odds Encoding

The Bayesian model can accommodate a general mapping F and we will be able to infer it from data. Here, we justify the popular choice of a log-odds mapping (as assumed by Khaw et al. [17, 18], Zhang et al. [30]), and how this leads to a reinterpretation of BLO as an approximation of the Bayesian model. Suppose that observers encode positive counts (κ_+) and negative counts (κ_-):

$$m=\begin{pmatrix}F_{num}(\kappa_+)+\epsilon_1\\ F_{num}(\kappa_-)+\epsilon_2\end{pmatrix}\qquad \epsilon_1,\epsilon_2\sim\mathcal{N}(0,\sigma^2) \tag{32}$$

where $p=\frac{\kappa_+}{\kappa_++\kappa_-}$. Based on research on magnitude perception, we take F_{num} to be consistent with Weber's law by assuming the form [e.g. 20, 4]:

$$F_{num}(\kappa)=\log(\kappa+\alpha)\qquad\text{hence}\qquad m=\begin{pmatrix}\log(pN+\alpha)+\epsilon_1\\ \log((1-p)N+\alpha)+\epsilon_2\end{pmatrix}$$

where $\alpha>0$ prevents an infinite FI at zero [20], and $N=\kappa_++\kappa_-$. What is an optimal 1D encoding $m_{1D}\in\mathbb{R}$ of the rate $p=\frac{\kappa_+}{\kappa_++\kappa_-}$? We focus on linear encodings with coefficients w_1, w_2:

$$m_{1D} := w_1(\log(pN+\alpha)+\epsilon_1)+w_2(\log((1-p)N+\alpha)+\epsilon_2) \tag{33}$$

This uniquely encodes p only if $w_1w_2<0$ because of monotonicity. Symmetry thus suggests $w_1=-w_2$, equivalent to a log-odds encoding smoothed at the boundaries ($\beta>0$):

$$F(p) := \log\frac{p+\beta}{(1-p)+\beta}\qquad \sqrt{\mathcal{J}(p)}=\frac{1}{\sigma}\cdot\left(\frac{1}{p+\beta}+\frac{1}{(1-p)+\beta}\right) \tag{34}$$

where, for fixed N, we absorbed N into $\beta=\alpha/N$. Note that this $\mathcal{J}(p)$ is U-shaped as in Figure 2A. This form for F allows us to interpret BLO as an approximation to the Bayesian model. First, in the limit where $\beta\to 0$ (i.e., unbounded log-odds encoding), the Bayesian bias comes out to (see Corollary 35 in Appendix B.1.1):

$$\mathbb{E}[\hat{p}|p]-p=\underbrace{\sigma^2p^2(1-p)^2\left(\log P_{prior}(p)\right)'}_{\text{Attraction}}+\underbrace{2\sigma^2p(1-p)(1-2p)}_{\text{Repulsion}}+\mathcal{O}(\sigma^4) \tag{35}$$

The first term depends on the prior distribution. As expected, the second term describes an S-shaped bias (Figure 2B). Indeed, when noise parameters are small, the BLO model matches the bias of this Bayesian model with a specific unimodal prior (Proof in Appendix B.1.2):

Theorem 10. *The BLO model in the limit of untruncated log-odds, and a Bayesian model with unbounded log-odds encoding ($\beta = 0$) and a specific unimodal prior $P_{prior}(p)$ (depending on Λ_0) have the same bias up to difference $\mathcal{O}(\kappa^2 + \sigma^4)$.*

This result suggests that one can reinterpret BLO as an approximation of a Bayesian model with log-odds encoding and a unimodal prior, albeit with divergences in the response distribution due to its sub-optimal decoding process. Our empirical results detailed below support this view, showing that the Bayesian model consistently provides better fit to behavior compared to the BLO model.

B.3 FI for Efficient Code in Frydman and Jin [10]

Frydman and Jin [10] take the encoding to be given by the code derived in Heng et al. [14], which is defined via:

$$\theta(p) = \sin^2\left(\frac{\pi}{2}F(p)\right) \tag{36}$$

where $F(p)$ is the cumulative distribution function of the prior, whose density we denote $f(p)$. We now note, for $a(p) := \frac{\pi}{2}F(p)$:

$$\theta'(p) = 2\sin a \cos a \cdot a'(p) = \sin(2a) \cdot a'(p) \tag{37}$$

and hence $a'(p) = \frac{\pi}{2}f(p)$. Now for a code given by $m \sim \text{Binomial}(n, \theta(p))$, the Fisher Information is

$$\begin{aligned}
\mathcal{J}(p) =& \frac{n[\theta'(p)]^2}{\theta(p)(1-\theta(p))} \\
=& \frac{n\sin^2(2a)[a'(p)]^2}{\sin^2(a)(1-\sin^2(a))} \\
=& \frac{n\sin^2(2a)[a'(p)]^2}{\sin^2(a)\cos^2(a)} \\
=& \frac{n\sin^2(2a)[a'(p)]^2}{\frac{1}{4}\sin^2(2a)} \\
=& 4n[a'(p)]^2 \\
=& 4n(\frac{\pi}{2}f(p))^2 \\
\propto& f(p)^2
\end{aligned}$$

We thus obtain $\sqrt{\mathcal{J}(p)} \propto p_{prior}(p)$.

C Experimental Methods and Supplementary Results

C.1 Theoretical Mapping of Model Variants

Table 1: Correspondence between existing theoretical accounts and the model variants used in our empirical comparisons.

Theoretical account	Corresponding model variant
Fixed probability distortion functions	LILO, Prelec
Log-odds-based accounts	BLO, Bayesian model with bounded log-odds encoding
Regression-based accounts	Bayesian model with uniform encoding
Efficient-coding account	Bayesian model with prior-matched encoding
General Bayesian account	Bayesian model with freely fitted prior and encoding

C.2 Task and Datasets

C.2.1 Judgment of Relative Frequency (JRF)

We analyzed the publicly available dataset from Zhang et al. [30] and Zhang and Maloney [29] . On each trial, subjects were briefly shown an array of black and white dots and reported their estimate of the relative frequency of one color by clicking on a horizontal scale.

We used two datasets that differ in the granularity of the stimulus proportions:

- **Zhang et al. [30] Dataset (JD, including JDA and JDB):**
  - **Subjects:** A total of 75 subjects were divided into two groups: 51 subjects who completed 660 trials (JDA) and 24 who completed 330 trials (JDB).
  - **Stimuli:** Coarse-grained. The objective relative frequency was drawn from the 11 discrete probability levels: 0.01, 0.05, 0.1, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9, 0.95, 0.99.
- **Zhang and Maloney [29] Dataset (ZM12):**
  - **Subjects:** 11 subjects who completed 800 trials.
  - **Stimuli:** Fine-grained. Stimulus proportions consisted of 99 levels, ranging from 0.01 to 0.99.

In both dataset, the total number of dots on any given trial was one of five values: 200, 300, 400, 500, or 600. For all datasets, each trial included the true stimulus proportion, the subject's estimate, total number of dots and other information, such as reaction time, that we didn't use for model fitting. We did not apply any data preprocessing.

C.2.2 Pricing Task in Decision-making Under Risk (DMR)

For this task, we analyzed the publicly available dataset from Zhang et al. [30]. This task used the same procedure and design as Gonzalez and Wu [11]. On each trial, subjects were presented with a two-outcome monetary gamble (e.g., a 50% chance to win \$100 or \$0 otherwise) and a table of sure amounts. subjects made a series of choices between the gamble and the sure amounts, a process that used a sequential bisection method to narrow the range and determine their Certainty Equivalent (CE) for the gamble.

- **Subjects:** The dataset comprises responses from the same 75 subjects who participated in the JD dataset in JRF task described above. 51 subjects from JDA performed 330 trials each and 24 subjects from JDB performed 165 trials each.
- **Stimuli:** The experiment used 15 distinct pairs of **non-negative** outcomes (e.g., \$25 vs. \$0; \$100 vs. \$50; \$800 vs. \$0). These were crossed with the same 11 probability levels from the JRF task for the higher outcome, resulting in 165 unique gambles.
- **Recorded Data:** Each recorded data point included the gamble's two outcomes, the probability of the higher outcome, the subject's final determined CE and other information that we didn't use for model fitting.

We did not apply any data preprocessing.

C.2.3 Choice Task in Decision-making Under Risk (DMR)

For this task, we analyzed the publicly available dataset from the Choice Prediction Competition 2015 (CPC15) [6]. This study was designed to test and compare models on their ability to predict choices between gambles that elicit classic decision-making anomalies. On each trial, subjects were presented with two or more distinct monetary gambles and made a one-shot choice indicating their preference.

The full CPC15 dataset contains a wide range of gamble types, including ambiguous gambles (Amb=1), gambles with correlated outcomes (Corr $\neq 0$), and multi-outcome lotteries (LotNum > 1). For the purpose of our analysis, which requires two-outcome gambles with fully specified probabilities and independent payoffs, we applied the following filters:

- Amb = 0: excluded ambiguous gambles, i.e., those with probabilities not explicitly described to subjects.
- Corr = 0: excluded gambles with correlated payoffs across options.
- LotNum = 1: restricted to gambles with exactly two outcomes in each option.

This gives us a subset of the CPC15 dataset with following information:

- **Subjects:** The subset contains responses from 153 subjects from the competition's estimation and test sets.
- **Stimuli:** The gambles covered 14 different behavioral phenomena, including the Allais and Ellsberg paradoxes. Unlike the pricing task, these gambles included both **gains and losses**, and the probabilities were drawn from a larger set of levels.
- **Data:** Each data point recorded the two gambles presented, the subject's choice, and information we didn't use for model fitting.

C.3 Adaptation Experiment

This experiment used the JRF dot-counting paradigm, but critically, the distribution of stimulus proportions was manipulated to be bimodal to test the model's ability to adapt its prior. Data was collected on an online platform following the procedure in Zhang et al. [30].

The bimodal dataset comprises responses from **26 subjects** across several designs to assess adaptation:

- 5 subjects performed 740 trials following from a bimodal distribution.
- 13 subjects performed an initial 740 trials following a bimodal distribution, followed by uniform trials to assess after-effects. In the 13 subjects, 7 subjects performed 840 trials(740 bimodal trials followed by 100 uniform trials), and 6 subjects performed 1136 trials(740 bimodal trials followed by 396 uniform trials).
- 8 subjects performed 942 bimodal trials interspersed with 198 uniform trials.

For all subjects in our collected datasets, in addition to the true number of black and white dots, the color designated for estimation, the subject's estimated proportion, and the reaction time, we also varies and logged the display time for each trial. We excluded estimated proportions of exactly 0% or 100% from the analysis.

C.3.1 Recruitment and Task instructions

Recruitment and compensation. Participants were recruited online via Prolific (`prolific.co`), with eligibility limited to adults (18+) with normal or corrected-to-normal vision. We did not collect further demographic information, including participants' exact ages, gender, country of residence, or language background. Recruitment complied with Prolific's policies. The experiment was expected to take less than 120 minutes; payment was prorated by expected completion time to exceed the German minimum wage. Pay and time estimates were adjusted during recruitment if observed completion time deviated from expectation. Subject identities were stored as anonymized numerical identifiers (subject 1, subject 2, ...).

Task instructions. Participants were shown the following clarification at the start of the experiment:

> "Thank you for participating in this study. In this experiment, you will be estimating the relative frequency of visual elements (black and white dots). You must be at least 18 years of age and legally able to consent to your participation in order to take part in this study.
>
> You will be presented with a display of black and white dots. The display will disappear after a short duration. After viewing the display, you will be asked to estimate the proportion of the specified color of dots by selecting a value on a horizontal bar ranging from 0% to 100%.
>
> The entire experiment is expected to take less than 120 minutes. If you are ready to begin, please turn off any distractions such as music or television, and click below to begin. Thank you!"

Informed consent. Participants provided informed consent before beginning the study:

> "Your participation in this experiment is voluntary. You have the right to withdraw your consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. All data resulting from this study will be stored in anonymized form. The results of the experiment may, in anonymized form, be used and published for scientific purposes. By clicking on 'Continue', you declare that you agree to voluntarily participate in this experiment."

Optional post-task questionnaire. At the end of the experiment, participants were asked optional questions about (1) task comprehension, (2) study experience, (3) appropriate payment, and (4) free-text feedback. No personally identifiable information was collected.

Ethics approval. The experimental design was approved by the Ethical Review Board of the Faculty of Mathematics and Computer Science, Saarland University.

C.4 Model Fitting and Evaluation

C.4.1 Implementation and Optimization Details

Across all datasets, we ran all Bayesian models using the implementation from Hahn and Wei [12], Hahn et al. [13], using a grid of 200 points and regularization strength 1.0. To evaluate sensitivity to these hyperparameters, we additionally refitted FreeP+FreeE using alternative grid sizes and regularization strengths (Appendix C.5.6).

For all the Bayesian models, as well as the reimplemented Zhang et al. [30]'s models, the optimization was performed using a gradient-based approach(Adam or SignSGD). All fits run on a single consumer-grade GPU (e.g., NVIDIA Titan X) and are also feasible on CPU. Per-subject fit time varies by dataset and model variant, ranging from minutes to hours. Total compute across all paradigms (JRF, lottery pricing, risky choice, bimodal adaptation) and model variants is on the order of GPU-days. Fits are independent across subjects and trivially parallelizable across compute nodes. Reproduction does not require specialized compute infrastructure.

C.4.2 Model Comparison Metrics

We primarily use summed held-out ΔNLL to evaluate predictive performance, complemented by ΔAICc analyses based on estimated effective degrees of freedom.

Summed Heldout ΔNLL. This metric measures a model's generalization performance, without penalizing for model complexity. The procedure of using this metric is as follows:

1. For each subject, the data are partitioned into ten folds. We use fold 0 as the held-out test set and the remaining nine folds as the training set. The same split is used for all model variants.
2. The fitted model is evaluated by calculating its negative log-likelihood (NLL) on the held-out test set. A lower NLL indicates better predictive performance.
3. For each subject, ΔNLL is a given model's held-out NLL minus the lowest held-out NLL across the model set. The summed held-out ΔNLL is the sum of these values across all subjects.

Summed ΔAICc. Following Zhang et al. [30], we use AICc as a complementary model comparison metric, adjusting model complexity using estimated effective degrees of freedom for regularized nonparametric models:

1. For each subject, a model is fitted on all trials to obtain its NLL. We use the estimated effective degrees of freedom k_{eff} for regularized nonparametric models (as described below) and the number of identifiable parameters for fully parametric models. We denote either complexity measure by k.
2. We compute the effective-df-adjusted corrected Akaike Information Criterion:

$$\text{AICc} = 2\,\text{NLL} + 2k + \frac{2k(k+1)}{n-k-1}, \tag{38}$$

where n is the number of trials per subject and $n > k + 1$ is required.

3. For each subject, $\Delta\mathrm{AICc}$ is a given model's AICc minus the lowest AICc across the model set. The summed $\Delta\mathrm{AICc}$ is the sum of these values across the retained participants.

Effective degrees of freedom for regularized models. For smoothly regularized nonparametric models, counting every grid point as a free parameter overestimates model complexity. We estimate the local effective degrees of freedom at the fitted regularization strength as

$$k_{\mathrm{eff}} = \mathrm{tr}\left[(\mathbf{F} + \mathbf{S})^{+}\mathbf{F}\right], \tag{39}$$

where $\mathbf{F}$ is the empirical Fisher information computed from the task-specific pointwise likelihood, $\mathbf{S}$ is the Hessian of the smoothness penalty, and $+$ denotes the Moore–Penrose pseudoinverse, which accommodates numerically flat directions in the overparameterized representation.

Both matrices are computed over the complete identifiable parameter vector, including unregularized scalar parameters. These parameters have zero penalty curvature and contribute approximately one degree of freedom per identifiable direction. Fully parametric models retain their identifiable parameter counts. We recompute the empirical Fisher information for each likelihood specification (analytical, categorical, or binary-choice) and use k_{eff} in place of the nominal parameter count in the participant-level AICc calculation.

Because this curvature-based approximation is local, we restrict the analysis to fits sufficiently close to a stationary point of the complete penalized objective,

$$\mathcal{L}_{\mathrm{pen}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\ell_i(\theta) + \mathrm{Penalty}(\theta). \tag{40}$$

Specifically, we retain fits satisfying

$$\left\|\nabla_\theta \mathcal{L}_{\mathrm{pen}}\right\|_\infty \leq 0.01. \tag{41}$$

For each task and likelihood specification, model comparisons are performed on the common subset of participants satisfying this criterion for all regularized models being compared. We report the retained sample sizes alongside the comparison results.

We do not report effective-df-adjusted AICc for the DMR pricing task because its two-stage fitting procedure does not provide a well-defined joint complexity estimate for the final certainty-equivalent likelihood. Model comparisons for this task are based on held-out NLL.

C.4.3 Categorical and Analytical Fitting

We use two ways to model motor noise in participants' responses. The analytical version treats responses as continuous and computes the likelihood using a continuous probability density. The categorical version discretizes the response space into a grid and computes likelihoods using the corresponding density approximations.

The two ways are motivated by the experimental procedure to obtain the responses. In JRF and adaptation, participants report probabilities using a continuous response bar, so we use the analytical version as the default specification. In DMR pricing, certainty equivalents (CEs) are obtained through a sequence of discrete choices between a gamble and sure amounts, so we use the categorical version as the default specification. We also evaluate both versions for each of these tasks to check whether the results depend on the likelihood specification.

JRF and Adaptation. Analytical Version. We treat responses as continuous and assume they are drawn from a Gaussian centered on the Bayesian estimate $\hat{p}(m)$ with variance σ^2_{motor}. Because responses are bounded by the grid $[r_{\min}, r_{\max}]$, the Gaussian is truncated and normalized using the corresponding CDF values. This gives the exact continuous likelihood, though it can be numerically unstable when the motor variance is very small. Finally, we mix this motor likelihood with a uniform component to account for guessing:

$$
\begin{aligned}
Z &= \Phi\left(\tfrac{p_{\max}-\hat{p}(m)}{\sigma_{\text{motor}}}\right) - \Phi\left(\tfrac{p_{\min}-\hat{p}(m)}{\sigma_{\text{motor}}}\right) \\
P(p_{\text{obs}} \mid m) &= \frac{1}{Z}\frac{1}{\sqrt{2\pi\sigma^2_{\text{motor}}}} \exp\left(-\tfrac{(p_{\text{obs}}-\hat{p}(m))^2}{2\sigma^2_{\text{motor}}}\right) \\
P_{\text{mix}}(p_{\text{obs}} \mid m) &= (1-u)\cdot P(p_{\text{obs}} \mid m) + u.
\end{aligned}
$$

$$
\begin{aligned}
Z_m &= \Phi\left(\frac{p_{\max}-\hat{p}(m)}{\sigma_{\text{motor}}}\right) - \Phi\left(\frac{p_{\min}-\hat{p}(m)}{\sigma_{\text{motor}}}\right), \\
f(p_{\text{obs}} \mid m) &= \frac{1}{Z_m\sqrt{2\pi\sigma^2_{\text{motor}}}} \exp\left(-\frac{(p_{\text{obs}}-\hat{p}(m))^2}{2\sigma^2_{\text{motor}}}\right), \\
f_{\text{mix}}(p_{\text{obs}} \mid m) &= (1-u)f(p_{\text{obs}} \mid m) + u.
\end{aligned}
$$

Categorical Version. We discretize responses into $J = 200$ equally spaced grid points $\{p_j\}_{j=1}^{J}$, using the same grid as for the input stimuli. We assign a Gaussian probability mass to each grid point, centered on the Bayesian estimate $\hat{p}(m)$ with variance σ^2_{motor}, and normalize these probabilities using log-softmax. Each observed response is assigned to the nearest grid point. We then mix the corresponding probability mass with the uniform guessing component:

$$
\begin{aligned}
\log q_j(m) &= \text{logsoftmax}_j\left(-\frac{(p_j-\hat{p}(m))^2}{2\sigma^2_{\text{motor}}}\right), \\
j_{\text{obs}} &= \arg\min_j |p_{\text{obs}} - p_j|, \\
P_{\text{mix}}(j_{\text{obs}} \mid m) &= (1-u)q_{j_{\text{obs}}}(m) + \frac{u}{J}.
\end{aligned}
$$

DMR Pricing. Analytical Version. We treat the CE report as a continuous variable and assume it is drawn from a Gaussian centered on the model prediction μ_m with variance σ^2_{motor}, truncated to the response interval $[x_2, x_1]$ allowed on each trial. This provides the exact continuous likelihood, though it can become numerically unstable when σ^2_{motor} is very small. We mix this motor likelihood with a continuous uniform distribution to account for guessing, where $x_1 > x_2$ the two lottery outcomes on the current trial:

$$
\begin{aligned}
Z_m &= \Phi\left(\frac{x_1-\mu_m}{\sigma_{\text{motor}}}\right) - \Phi\left(\frac{x_2-\mu_m}{\sigma_{\text{motor}}}\right), \\
f(\text{CE}_{\text{obs}} \mid m) &= \frac{1}{Z_m\sqrt{2\pi\sigma^2_{\text{motor}}}} \exp\left(-\frac{(\text{CE}_{\text{obs}}-\mu_m)^2}{2\sigma^2_{\text{motor}}}\right), \\
f_{\text{mix}}(\text{CE}_{\text{obs}} \mid m) &= (1-u)f(\text{CE}_{\text{obs}} \mid m) + \frac{u}{x_1-x_2}.
\end{aligned}
$$

Categorical Version. We discretize the CE response axis into $J = 1000$ equally spaced grid points $\{c_j\}_{j=1}^{J}$. With $\text{CE}_{\min} = 0$ and $\text{CE}_{\max} = 800$, the grid spacing is $\Delta c = (\text{CE}_{\max} - \text{CE}_{\min})/J = 0.8$. The grid points are therefore $c_j = (j-1)\Delta c$, ranging from 0 to 799.2; the upper endpoint 800 is excluded. We normalize Gaussian scores over the CE grid to obtain the probability mass $q_j(m)$ assigned to each grid point. Then, for an observed response CE_{obs}, we select its nearest grid point and convert the selected probability mass into a density approximation. We mix the motor density approximation with a uniform guessing density between these outcomes:

$$
\begin{aligned}
\log q_j(m) &= \text{logsoftmax}_j\left(-\frac{(c_j-\mu_m)^2}{2\sigma^2_{\text{motor}}}\right), \\
j_{\text{obs}} &= \arg\min_j |\text{CE}_{\text{obs}} - c_j|, \\
P_{\text{mix}}(\text{CE}_{\text{obs}} \mid m) &= (1-u)p(\text{CE}_{\text{obs}} \mid m) + \frac{u}{x_1-x_2}.
\end{aligned}
$$

C.5 Details on Bayesian Model for JRF Task

C.5.1 Bayesian Model Variants

We tested several variants of our Bayesian framework by combining different priors and encodings; some are discussed in Section 4.1:

- **Priors:**
  - **Uniform Prior (UniformP):** This variant assumes a uniform distribution of prior over the range of possible stimuli(i.e., across all grid points). There is no learnable parameter.
  - **Gaussian Prior (GaussianP):** This variant assumes a Gaussian distribution over the range of possible grid points. The two fitted parameters are Gaussian mean and Gaussian standard deviation.
  - **Freely Fitted Prior (FreeP):** In this variant, all values of the prior distribution across 200 grid points are treated as trainable parameters. We estimate the local effective degrees of freedom at the fitted regularization strength, see Appendix C.4.2.

- **Encodings**
  - **Uniform Encoding (UniformE):** This encoding assumes a uniform distribution of encoding over the range of stimuli. There is no fitted parameter.
  - **Fixed Unbounded Log-Odds Encoding (UnboundedLOE):** The encoding is proportional to the log-odds of the stimuli(grid values). This is consistent with the log-odds assumption underlying Zhang and Maloney [29]'s unbounded log-odds models. There is no fitted parameter.
  - **Bounded Log-Odds Encoding (BoundedLOE):** This encoding is given by $F'(p) = \frac{1}{(p+\beta_1)(1-p+\beta_2)}$. Note that this is equivalent to (34) in the setting where different β's, written here as β_1, β_2, are allowed for positive and negative counts, up to an (irrelevant, as it doesn't depend on p) proportionality constant:

    $$\frac{1}{\sigma}\left(\frac{1}{p+\beta_1}+\frac{1}{(1-p)+\beta_2}\right) \propto \frac{1+\beta_2+\beta_1}{(p+\beta_1)(1-p+\beta_2)} \propto \frac{1}{(p+\beta_1)(1-p+\beta_2)} \tag{42}$$

    , where x is the value of grid, and β_1 and β_2 are small, positive, learnable parameters between 0 and 1, which we bound using the sigmoid function. This form is motivated by Zhang's BLO models and our discussion in Section B.2. There are two fitted parameters: β_1 and β_2.
  - **Prior-matched Encoding(PriorMatchedE):** This encoding assumes that the resources is proportional to the prior distribution, i.e., the encoding density is identical to the prior. There are no additional fitted parameters beyond those of the prior.
  - **Freely Fitted Encoding (FreeE):** Similar to freely fitted prior. There are 200 freely fitted parameters.

Parameters for each Bayesian model variant, including prior parameters, encoding parameters, sensory noise variance, motor variance and mixture logit are optimized against the subject-level data using the same gradient based method described for BLO models on the same task.

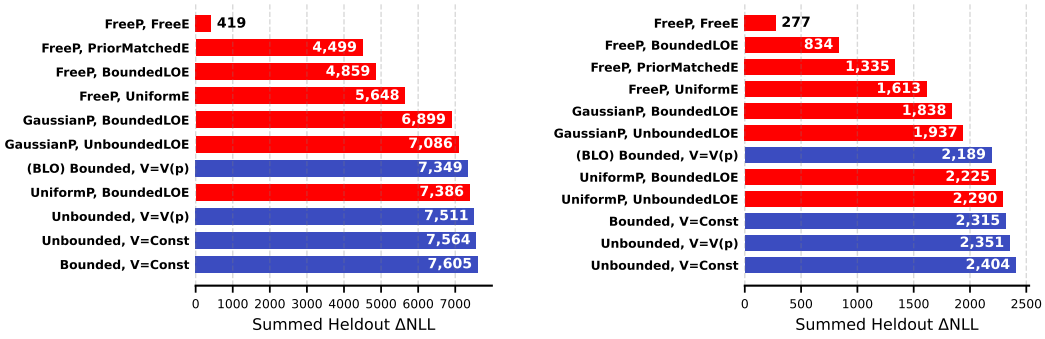
C.5.2 Performance Comparison of All Model Variants

We compare all model variants using summed held-out ΔNLL and effective-df-adjusted ΔAICc under both analytical and categorical likelihood specifications (Figures 8 and 9). Bayesian model variants (red) are described in Appendix C.5.1, while variants from Zhang et al. [30] (blue) are described in Appendix C.10.1.

FreeP+FreeE achieves the lowest summed held-out ΔNLL under both likelihood specifications, with predictive performance evaluated across all 86 participants. The complementary effective-df-adjusted AICc analysis also favors FreeP+FreeE, although it is restricted to participants whose model fits satisfy the stationarity criterion (Appendix C.4.2). This retains 75 of 86 participants under the analytical likelihood and 74 of 86 under the categorical likelihood. The median effective degrees

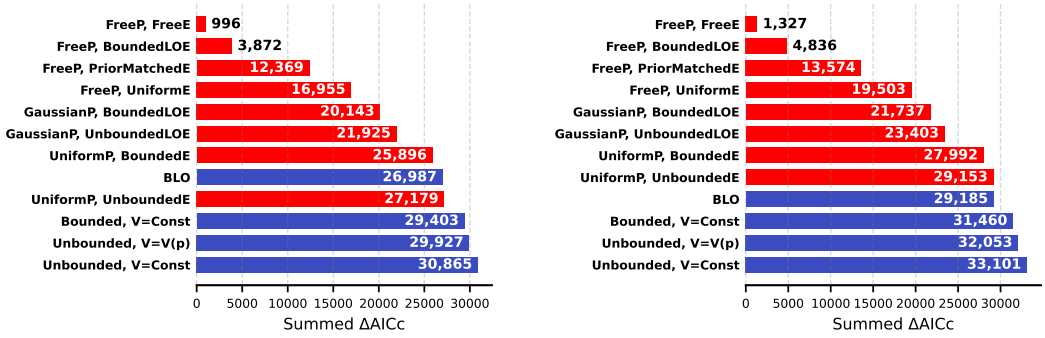
of freedom for FreeP+FreeE are 52.9 [IQR: 44.4, 65.7] and 54.9 [IQR: 45.0, 66.5] respectively. FreeP+FreeE achieves the lowest participant-level AICc for 57/75 (76%) and 60/74 (81%) of the retained participants.

Overall, our model performance better than Zhang et al. [30]'s model variants across two metrics and two fitting approaches. Within Bayesian models, models using a bounded log-odds encoding outperform those with an unbounded encoding, and a Gaussian prior is superior to a uniform prior. Within Zhang et al. [30]'s model variants, bounded model performs better than unbounded model, and using $V(p)$(explained in Eq 47) is better than assuming a constant value V.



(A) Analytical (B) Categorical

Figure 8: **JRF Task:** Performance of the models, measured by the Summed Heldout ΔNLL metric. See Appendix C.5.1 for the shorthands for Bayesian models (red). We refer to Zhang et al. [30] for the shorthands for their model variants (blue). Analytical and categorical fitting methods are explained in Appendix C.4.3. The main-text results use the analytical likelihood.



(A) Analytical (B) Categorical

Figure 9: **JRF Task:** Performance of the models, measured by the Summed Effective-df-adjusted ΔAICc metric. Comparisons are restricted to the common subset of participants whose regularized model fits satisfy the stationarity criterion: 75/86 for the analytical likelihood (A) and 74/86 for the categorical likelihood (B). Effective degrees of freedom are estimated as described in Appendix C.4.2. The shorthands for Bayesian models (red) are in Appendix C.5.1, and we refer to Zhang et al. [30] for the shorthands for their model variants (blue). Analytical and categorical fitting methods are explained in Appendix C.4.3.

Sensitivity to the model comparison criterion. We additionally compared model performance using BIC with estimated effective degrees of freedom as a sensitivity analysis. As with the effective-df-adjusted AICc analysis, comparisons were restricted to participants whose regularized model fits satisfied the stationarity criterion required for effective-df estimation (Appendix C.4.2). This retained 75 of 86 participants under the analytical likelihood and 74 of 86 under the categorical likelihood.

Unlike held-out NLL and effective-df-adjusted AICc, the stronger $k \log n$ penalty in BIC amplifies the parameter count. The result favors FreeP+BoundedLOE over FreeP+FreeE under both likelihood specifications. FreeP+BoundedLOE achieves the lowest summed ΔBIC among all model variants, with values of 884 and 1,533 under the analytical and categorical likelihoods, compared with 9,105 and 9,113 for FreeP+FreeE.

C.5.3 Subject-level Held-out Model Comparisons

To examine whether the aggregate advantage of FreeP+FreeE is consistent across individual participants, we compared its held-out NLL with that of FreeP+UniformE, FreeP+PriorMatchedE, and BLO. For each participant s and competing model m, we computed the per-trial difference

$$\Delta_{s,m} = \frac{\mathrm{NLL}_{s,m} - \mathrm{NLL}_{s,\mathrm{FreeP+FreeE}}}{n_{s,\text{held-out}}}, \tag{43}$$

where positive values favor FreeP+FreeE. These comparisons use the same held-out test set (fold 0) and analytical likelihood as the aggregate model comparisons. We obtained 95% confidence intervals for the median paired differences by resampling participants with replacement 20,000 times.

Table 2: **JRF Task:** Subject-level held-out NLL comparisons. Positive per-trial differences favor FreeP+FreeE. Confidence intervals are obtained from 20,000 participant-level bootstrap samples.

Competing model	FreeP+FreeE wins	Median ΔNLL/trial [IQR]	95% CI
FreeP+UniformE	83/86 (96.5%)	0.243 [0.130, 0.379]	[0.200, 0.294]
FreeP+PriorMatchedE	79/86 (91.9%)	0.167 [0.064, 0.289]	[0.123, 0.207]
BLO	83/86 (96.5%)	0.235 [0.119, 0.631]	[0.205, 0.365]

As shown in Table 2, FreeP+FreeE achieves lower held-out NLL for 79–83 of the 86 subjects across the three comparisons. The median per-trial differences are positive in all cases, and all bootstrap confidence intervals exclude zero. These results show that the advantage of FreeP+FreeE is broadly consistent across subjects.

C.5.4 Analysis of Fitted Prior and Resources

Figure 10 plots the fitted resources for our Bayesian model (red) and the BLO model (blue) for 86 subjects in the JRF task. For nearly all subjects, the resources from both models are U-shaped, with peaks near the probability endpoints of $p = 0$ and $p = 1$.

A closer examination reveals a difference. As shown in Figure 11, which restricts the y-axis for clarity, the resources of the BLO model exhibit two points of discontinuity. This finding is formally predicted by and consistent with our Theorem 7.

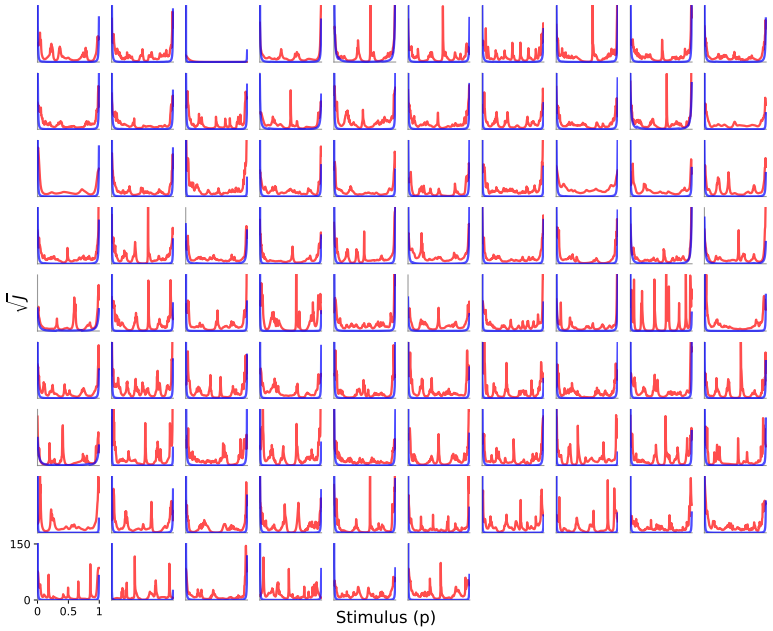


Figure 10: **JRF Task:** Per-subject resources ($\sqrt{J(p)}$) for the Bayesian model with freely fitted prior(in red) and encoding and the BLO model(in blue). Subjects S1 to S51 are from dataset JDA, S52 to S75 from dataset JDB (both from Zhang et al. [30]), and S76 to S86 from dataset ZM12 (Zhang and Maloney [29]).

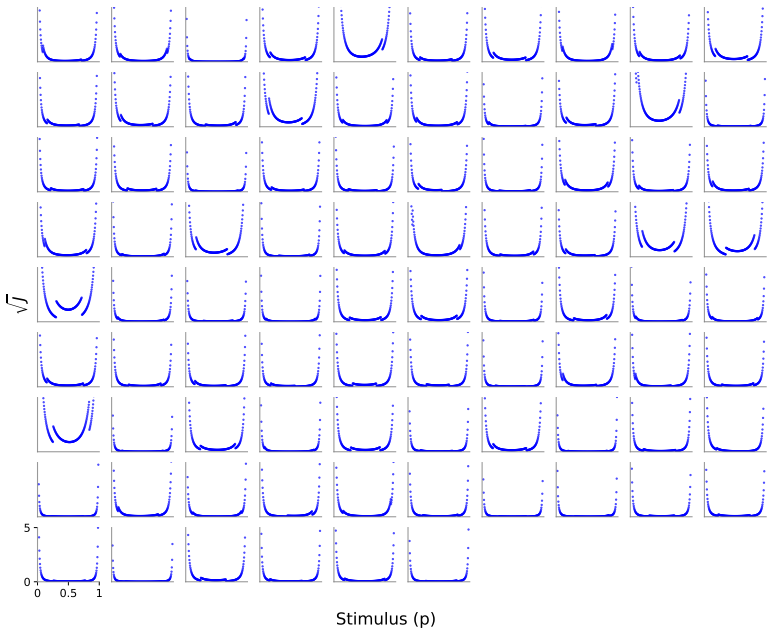


Figure 11: **JRF Task:** Discontinuity observed in the resources ($\sqrt{J(p)}$) of the BLO model. The proof of this property is in Theorem 7. Subjects S1 to S51 are from dataset JDA, S52 to S75 from dataset JDB (both from Zhang et al. [30]), and S76 to S86 from dataset ZM12 (Zhang and Maloney [29]).

Figure 12 compares the group-level fitted priors for the Bayesian model variants. The parametric Gaussian prior is shown to capture the main features of the non-parametric, freely fitted prior. The freely fitted prior with matched encoding shows not only peaks at 0.5, but also at 0 and 1, which is a property of the Resources.

Figure 13 compares the group-level resources for all Bayesian model variants. The resources are U-shaped for the parametric models (the Bayesian log-odds variants), and the freely fitted resources also recover this U-shape.

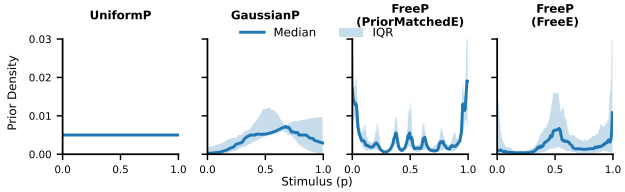


Figure 12: **JRF Task:** Group-level priors for the four prior components evaluated in this task. The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3.

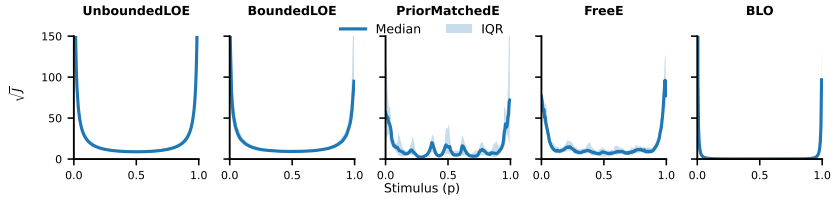


Figure 13: **JRF Task:** Group-level resources ($\sqrt{J(p)}$) for the four encoding components evaluated in this task, along with the resources of the BLO model. The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3. This result is mentioned in main text section 4.1.

C.5.5 Analysis of Bias and Variance

Figures 15 and 16 show the per-subject bias and variance, respectively, providing a more detailed view of the group-level results presented in Figure 14. The methods used to calculate these quantities are detailed in Appendix C.11.4.

The per-subject plots confirm the main findings. For bias, both the Bayesian model and the BLO model capture the bis pattern of the non-parametric estimates. The key divergence appears in the response variability. Figure 16 clearly shows that while our Bayesian model captures subject-level variability, the BLO model consistently fails to model the dip at $p = 0.5$.

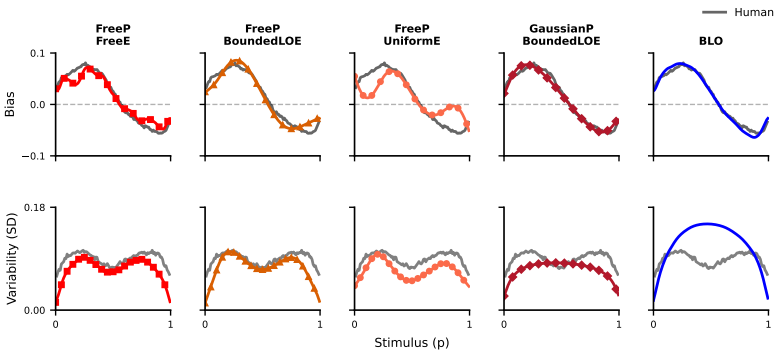


Figure 14: **JRF Task:** Model fits to response bias (top) and variability (bottom). Gray curves show human data. The S-shaped bias is only explained by non-uniform encodings (BoundedLOE, FreeE); the variability is only captured by the nonparametric prior (FreeP).

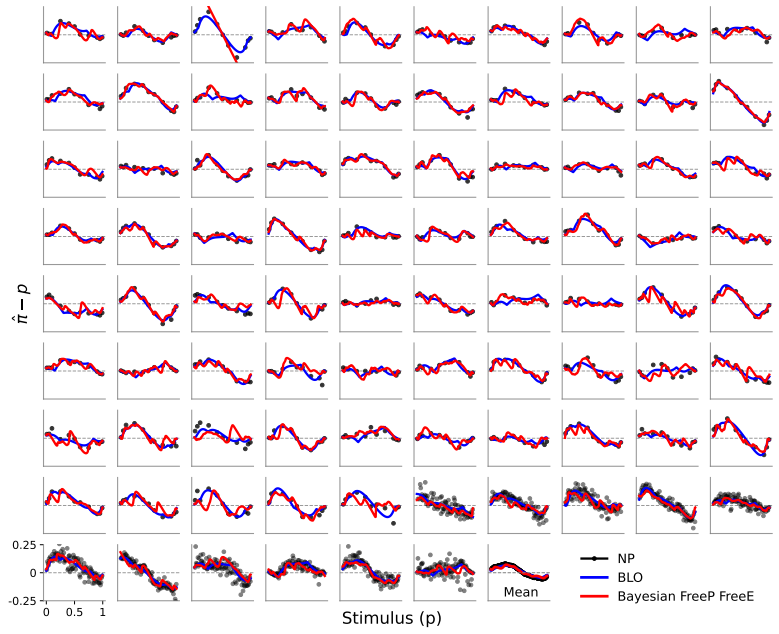


Figure 15: **JRF Task:** Per-subject bias of the non-parametric estimates from the data, the BLO model and the Bayesian model with free prior and free encoding. Subjects S1 to S51 are from dataset JDA, S52 to S75 from dataset JDB (both from Zhang et al. [30]), and S76 to S86 from dataset ZM12 (Zhang and Maloney [29]).

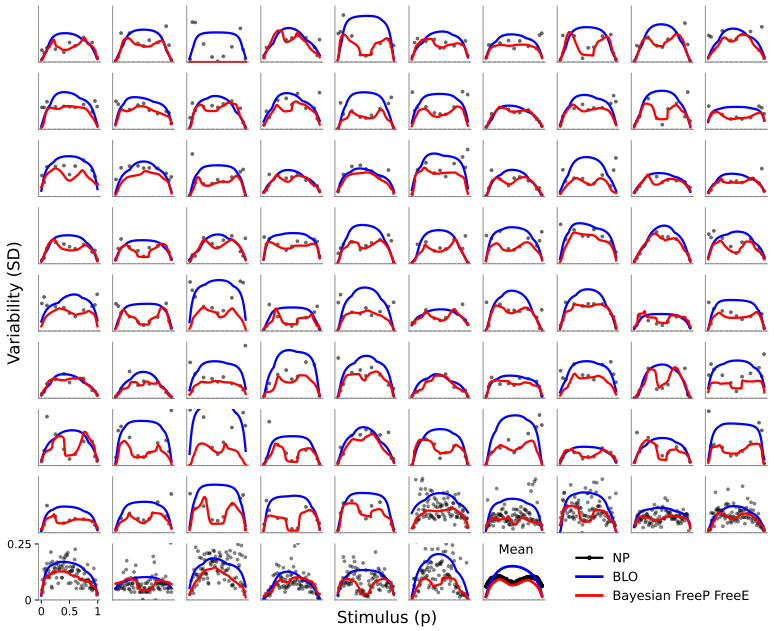


Figure 16: **JRF Task:** Per-subject variability of the non-parametric estimates from the data, the BLO model and the Bayesian model with free prior and free encoding. Subjects S1 to S51 are from dataset JDA, S52 to S75 from dataset JDB (both from Zhang et al. [30]), and S76 to S86 from dataset ZM12 (Zhang and Maloney [29]).

Moreover, Figure 40 decompose the bias of four Bayesian model variants into attraction, repulsion and boundary regression. Both attractive and repulsive components point away from 0 and 1. Both are needed to account for the overall distortion, as both prior and encoding need to be nonuniform to achieve good model fit.

C.5.6 Sensitivity to Grid Size, Regularization Strength, and Likelihood Specification

To assess the sensitivity of the recovered encoding to the fitting hyperparameters, we refitted the JRF FreeP+FreeE model using grid sizes of 200 and 400 points and regularization strengths of 0.1, 1.0, and 4.0, compared with the default setting of 200 points and regularization strength 1.0.

Across all six configurations, the group-level encoding resources retain U-shaped (Figure 17). Weaker regularization produces more local peaks, but the overall S-shaped pattern remains similar.

At the individual-subject level, the setting of 400 grid points and regularization strength 4.0 also recovers increased resources near the boundaries for many subjects, although individual are not uniformly U-shaped (Figure 18).

We also compared the recovered encoding under analytical and categorical likelihood specifications. Both yield similar group-level U-shaped resource profiles (Figure 19).

These results suggest that the group-level U-shaped encoding is not specific to the default fitting hyperparameters or likelihood specification.

C.5.7 Sensitivity to the Decoding Rule

Our main analyses assume posterior-mean decoding, which minimizes expected squared-error loss. To examine whether the recovered U-shaped encoding depends on this assumption, we refitted the JRF data using the posterior median, which minimizes expected absolute-error loss.

As shown in Figure 20, the group-level encoding resources remain U-shaped under both decoding rules, with higher resource allocation near the boundaries. At the individual-subject level, many participants also exhibit higher resources near the boundaries. (Figure 21).

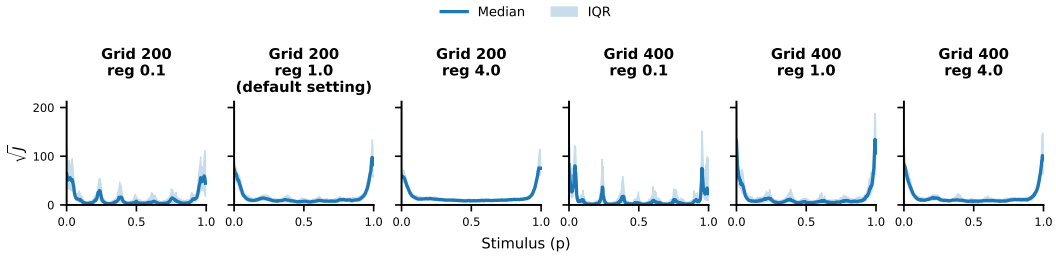


Figure 17: **JRF Task:** Sensitivity of recovered encoding resources ($\sqrt{\mathcal{J}(p)}$) to grid size and regularization strength. Group-level results are shown for grid sizes of 200 and 400 points and regularization strengths of 0.1, 1.0, and 4.0 across 86 participants. Solid lines indicate the median, and shaded regions show the interquartile range (IQR). All six configurations yield broadly U-shaped resource profiles.

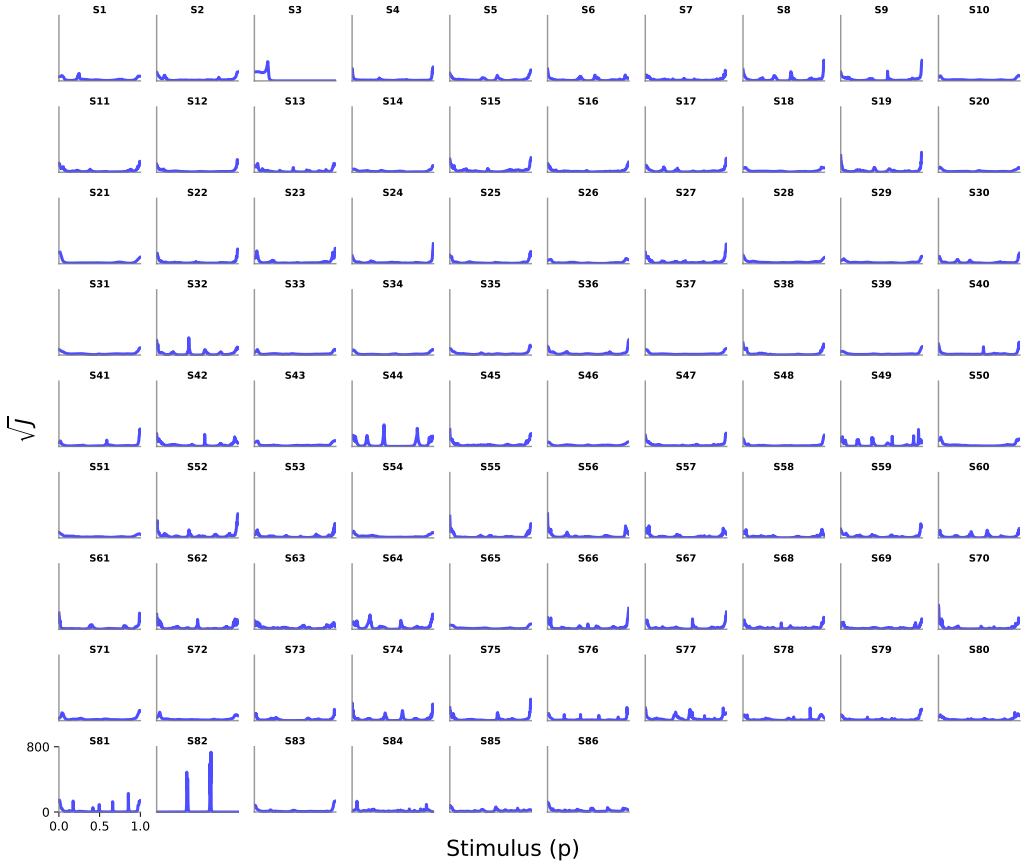


Figure 18: **JRF Task:** Individual-subject encoding resources ($\sqrt{\mathcal{J}(p)}$) recovered using 400 grid points and regularization strength 4.0 for all 86 participants. Many participants exhibit elevated resources near the probability boundaries, although individual profiles are heterogeneous.

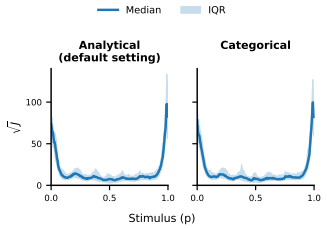


Figure 19: **JRF Task:** Group-level encoding resources ($\sqrt{\mathcal{J}(p)}$) recovered under analytical and categorical likelihood specifications, using the default grid size of 200 points and regularization strength 1.0. Solid lines indicate the median across 86 participants, and shaded regions show the interquartile range (IQR). Both specifications yield similar U-shaped profiles.

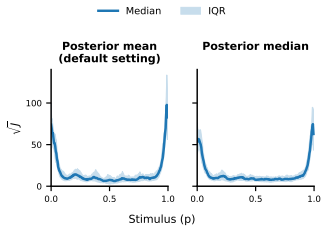


Figure 20: **JRF Task:** Group-level encoding resources ($\sqrt{\mathcal{J}(p)}$) under posterior-mean and posterior-median decoding. Solid lines indicate the median across 86 participants, and shaded regions show the interquartile range (IQR). Both decoding rules yield broadly U-shaped resource profiles.

C.5.8 Encoding Recovery Analysis

To examine whether the flexibility of FreeP+FreeE produces U-shaped encoding resources as an artifact, we conducte a recovery analysis. We generate synthetic JRF responses under three conditions: (i) uniform encoding with a uniform prior, (ii) uniform encoding with a bimodal prior, and (iii) U-shaped encoding with a bimodal prior. The generating noise parameters were set to the median values estimated from the 86 human participants fitted with FreeP+FreeE.

For each condition, we use 10 random seeds across three stimulus designs represented in the JRF data: 660 trials with 11 stimulus levels, 330 trials with 11 levels, and 800 trials with 99 levels. This gives 90 synthetic datasets, each refitted using the same FreeP+FreeE procedure applied to the human data. We classifie the recovered encoding by comparing its overlap coefficient with the uniform and U-shaped generating profiles, assigning it to whichever had the higher overlap. All recovered profiles are also inspected visually.

Table 3: **JRF Task:** Encoding recovery using FreeP+FreeE on 90 simulated datasets. Recovered encodings were classified by their overlap with the uniform and U-shaped generating profiles.

Generating encoding	Number of fits	Recovered UniformE	Recovered U-shapedE
UniformE	60	60	0
U-shapedE	30	0	30

As shown in Table 3, all 90 simulations were correctly classified. None of the 60 datasets generated with uniform encoding was recovered as U-shaped, even when the generating prior was bimodal.

C.6 Details on Bayesian Model for DMR Pricing Task

C.6.1 Bayesian Model Variants and Fitting Procdure

We tested the similar Bayesian model variants for the pricing task as we did for the JRF task (detailed in Appendix C.5.1). However, fitting these models to the Certainty Equivalent (CE) data required

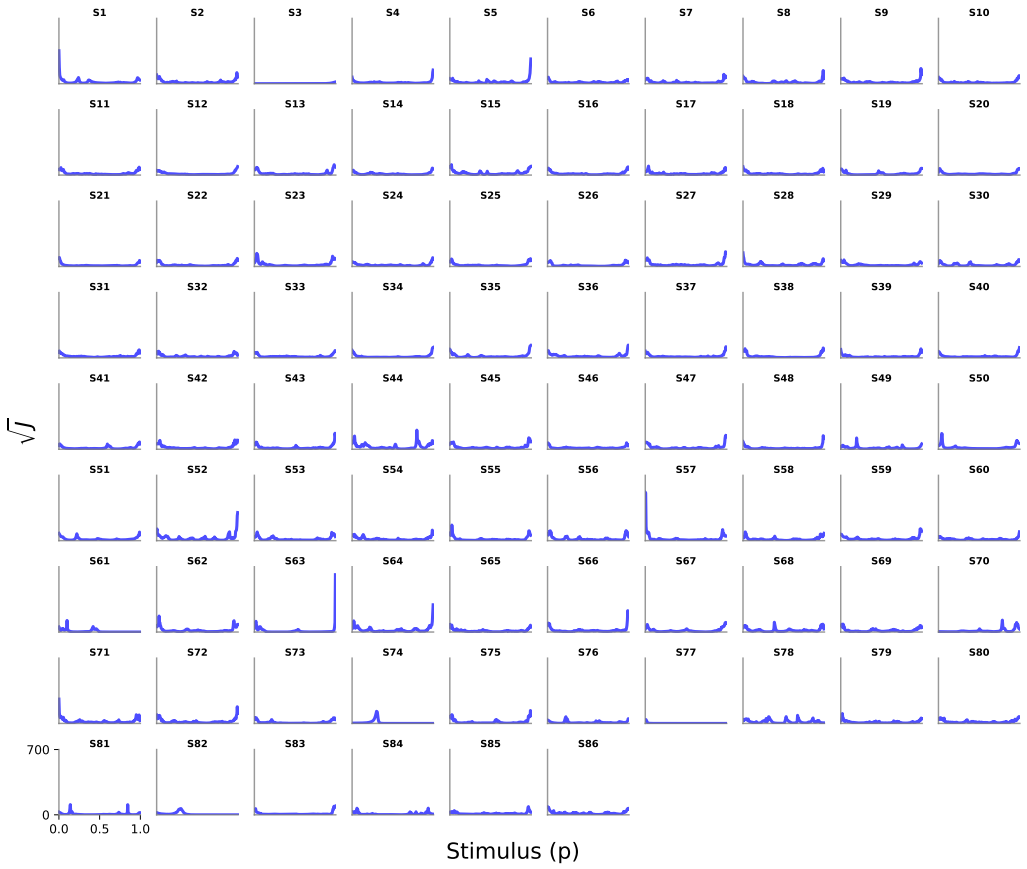


Figure 21: **JRF Task:** Individual-subject encoding resources ($\sqrt{\mathcal{J}(p)}$) recovered under posterior-median decoding for all 86 participants. The recovered resources generally exhibit higher resource allocation near the probability boundaries.

a specific two-stage procedure. The goal was to first convert each observed CE into an "implied" subjective probability weight, and then fit our Bayesian models to these weights.

Stage 1: Deriving trial-by-trial implied estimates. The goal of this initial stage was to convert each raw CE response into a non-parametric, trial-level estimate of the subjective probability weight, $\hat{\pi}_{\text{implied},t}(p)$. For each trial t, we fits a free variable. We also fit α for applying the CPT utility function. To account for additional variability in CE, we included an extra noise term ϵ_{CE} and optimized parameters by minimizing the loss between predicted and observed CE. This method is similar to that of Zhang et al. [30], but our implementation works on a trial-by-trial basis rather than on 11 discrete probability levels.

Stage 2: Fitting Bayesian estimator. The set of (p_t, $\hat{\pi}_{\text{implied},t}$) pairs derived from Stage 1 was then used as the target data to fit the parameters of our Bayesian model. These parameters in turn determine the set of optimal point estimates, $\hat{p}(m)$ (the decoded stimulus value for each possible internal measurement m), by maximizing the likelihood of the implied weights.

Stage 3: Final Likelihood Maximization on Original CE Data. To ensure a fair and direct comparison with the BLO model, the final model evaluation was based on the likelihood of the original CE data. In this final stage, the Bayesian encoding parameters (and therefore the set of $\theta(m)$ values) were held fixed from the results of Stage 2. We then performed a final optimization to find the subject's remaining parameters—the utility exponent α and the CE noise variance σ_{CE}—that maximized the log-likelihood of their observed CE responses. The resulting maximum log-likelihood value was then used to calculate the Held-out ΔNLL.

Model variants. Because stage 2 closely resembles the JRF task, we applied the same set of model variants used there (Appendix C.5.1), with the exception of the Fixed Unbounded Log-Odds Encoding. We excluded this variant to focus on encoding schemes that showed better performance in this task.

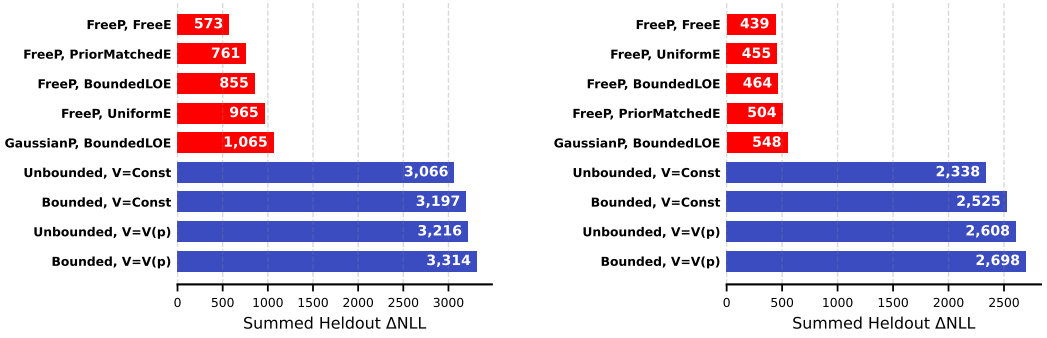
C.6.2 Performance Comparison of All Model Variants

Figure 22 compares the held-out predictive performance of all model variants under both analytical and categorical likelihood specifications.

For the likelihood of the observed CE data, we use the categorical specification in the main text, while reporting the analytical specification as a complementary analysis. To ensure a direct comparison, we re-evaluated the models of Zhang et al. [30] using the same categorical likelihood.

Under both likelihood specifications, FreeP+FreeE achieves the lowest aggregate held-out ΔNLL among the compared models. Subject-level comparisons further show a clear advantage over BLO, although the differences relative to FreeP+UniformE and FreeP+PriorMatchedE are inconclusive (Appendix C.4.2).

We do not report effective-df-adjusted AICc for this task because the two-stage fitting procedure does not provide a well-defined joint complexity estimate for the final CE likelihood, as mentioned in Appendix C.4.2.



(A) Analytical

(B) Categorical

Figure 22: **DMR Pricing Task:** Performance of the models, measured by the Heldout ΔNLL metric. See Appendix C.5.1 for the shorthands for Bayesian models (red). We refer to Zhang et al. [30] for the shorthands for their model variants (blue). Analytical and categorical fitting methods are explained in Appendix C.4.3. The results shown in the main text correspond to Categorical.

C.6.3 Subject-level Held-out Model Comparisons

To examine whether the aggregate predictive advantage of FreeP+FreeE is consistent across participants, we compared its held-out NLL with that of FreeP+UniformE, FreeP+PriorMatchedE, and BLO at the individual-subject level. For each participant and competing model, we computed the per-trial NLL difference relative to FreeP+FreeE, as described in Appendix C.5.3. Positive values favor FreeP+FreeE. We obtained 95% confidence intervals for the median paired differences by resampling participants with replacement 20,000 times.

Table 4: **DMR Pricing Task:** Subject-level held-out NLL comparisons across 75 participants. Positive per-trial differences favor FreeP+FreeE. Confidence intervals are obtained from 20,000 participant-level bootstrap samples.

Competing model	FreeP+FreeE wins	Median ΔNLL/trial	95% CI
FreeP+UniformE	40/75 (53.3%)	0.006	[-0.054, 0.056]
FreeP+PriorMatchedE	44/75 (58.7%)	0.048	[-0.007, 0.121]
BLO	71/75 (94.7%)	0.920	[0.737, 1.040]

As shown in Table 4, FreeP+FreeE achieves lower held-out NLL than BLO for 71 of the 75 participants, with a positive median per-trial difference and a bootstrap confidence interval excluding zero. In contrast, its advantages over FreeP+UniformE and FreeP+PriorMatchedE are less consistent across participants, with win rates slightly over 50%, and confidence intervals that include zero.

C.6.4 Analysis of Fitted Prior and Resources

The fitted resources in Figure 23 closely resemble those obtained in the JRF task. The BLO model doesn't have meaningful resources because it doesn't fit a noise in the encoding phase.

For the prior, the freely fitted version exhibits a shape similar to the Gaussian prior, which accounts for the strong performance of Bayesian models with Gaussian priors reported in Section C.6.2.

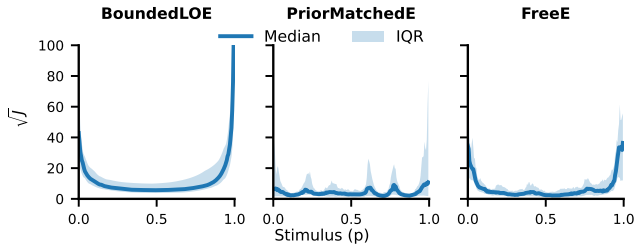


Figure 23: **DMR Pricing Task:** Group-level resources ($\sqrt{J(p)}$) for the three encoding components evaluated in this task. The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3.

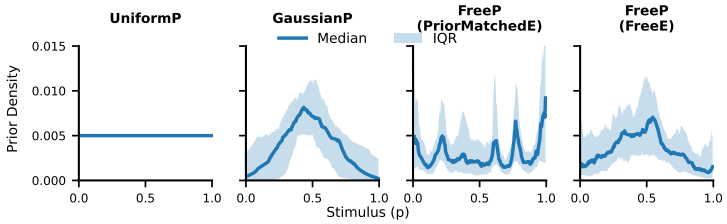


Figure 24: **DMR Pricing Task:** Group-level priors for the four prior components evaluated in this task. The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3. Free fitting is compatible with a unimodal prior.

C.6.5 Analysis of Bias and Variance

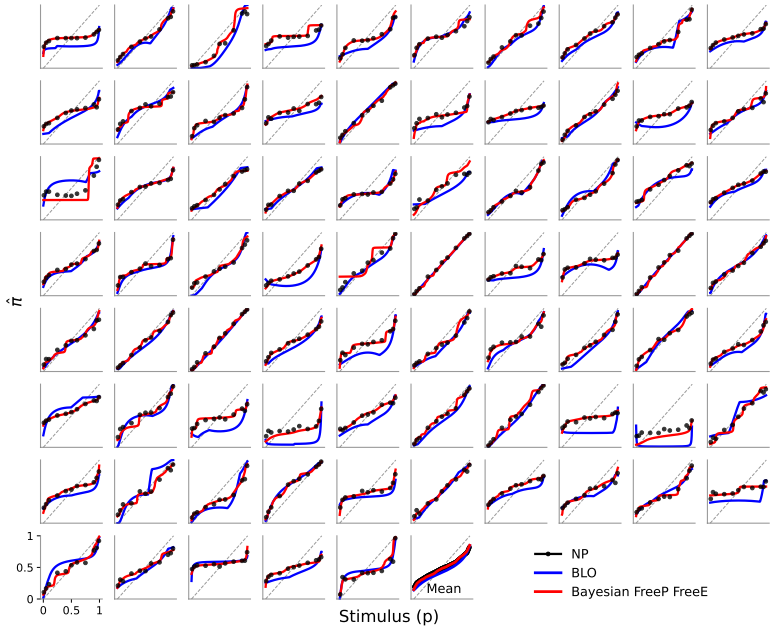


Figure 25: **DMR Pricing Task:** Per-subject bias of the non-parametric estimates from the data, the BLO model and the Bayesian model with free prior and free encoding on the DMR pricing task. Subjects S1 to S51 are from dataset JDA, and S52 to S75 are from dataset JDB (both from Zhang et al. [30]).

In Figure 25, we present the per-subject bias of the probability estimate, $\hat{\pi}$, for both the Bayesian and BLO models. As the figures show, both models capture the general pattern of bias for most subjects. A direct comparison of the variance of $\hat{\pi}$ is not shown because the two models treat this quantity fundamentally differently. The BLO model's estimate $\hat{\pi}_{\text{BLO}}$ is a deterministic point value and thus has zero variance. In contrast, our Bayesian estimate, $\hat{\pi}_{\text{Bayesian}}$, is the mean of a full posterior distribution and has inherent sensory noise variance. The methods used to calculate the non-parametric, BLO's, and Bayesian model's bias are detailed in Appendix C.11.4.

C.7 Details on Bayesian Model for DMR Choice Task

C.7.1 Full Utility Function, Logit Choice Rule, and LILO function

Full Utility Function

$$\text{Utility} = \begin{cases} w^+(p)v(X) + (1 - w^+(p))v(Y) & \text{if } X \geq Y \geq 0 \\ (1 - w^-(1-p))v(X) + w^-(1-p)v(Y) & \text{if } X > Y \text{ and } X, Y < 0 \\ w^+(p)v(X) + w^-(1-p)v(Y) & \text{if } X > 0 > Y \end{cases} \tag{44}$$

LILO function (γ, β are free parameters):

$$\widehat{p} = \lambda^{-1}(\gamma\lambda(p) + \beta), \quad \lambda(p) := \log \frac{p}{1-p}. \tag{45}$$

This equation is mentioned in main text section 4.2

Logit Choice Rule

$$P(\text{Choose B}) = \text{sigmoid}(\tau \cdot (\text{Utility}_B - \text{Utility}_A)) \tag{46}$$

The parameter τ ($\tau > 0$,) controls the choice sensitivity and is a free parameter to fit.

C.7.2 Bayesian Model Variants and Fitting Procedure

We tested several model variants:

- **Priors:**
  - **Freely Fitted Prior:** As we mainly focus on validating the resources shape with this task, we only evaluate the freely fitted prior, which is the same as described in Appendix C.5.1.
- **Encodings:**
  - **Bounded Log-odds Encoding, Prior-matched Encoding, Freely Fitted Encoding:** Same as in Appendix C.5.1.

Because the dataset includes both gains and losses, we fit separate probability weighting functions for them. Specifically, we estimate separate priors for gains and for losses, while assuming a shared encoding across both. The resulting probability weighting functions are then computed as the expectation of $\hat{p}$ under the encoding distribution.

C.7.3 Performance Comparison of Model Variants

Figure 26 shows the effective-df-adjusted ΔAICc comparison of Bayesian model variants in the DMR choice task. The analysis includes 152 of 153 participants whose fits satisfy the stationarity criterion (Appendix C.4.2). For FreeP+FreeE, the median effective degrees of freedom is 10.7 [IQR: 8.7, 12.2].

FreeP+FreeE achieves the lowest summed ΔAICc among the compared Bayesian variants and is the best-performing model for 121 of the 152 retained participants (80%). The best alternative is FreeP+BoundedLOE. These results provide complementary support for the advantage of FreeP+FreeE reported in Figure 5 in the main text.

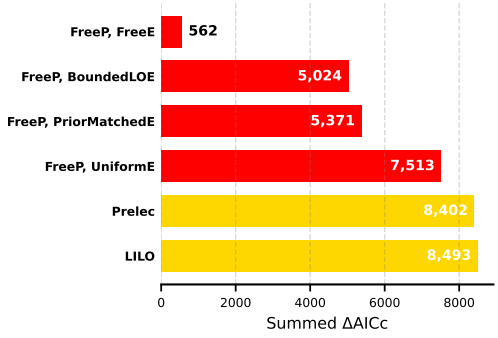


Figure 26: **DMR Choice Task:** Model comparison using summed effective-df-adjusted ΔAICc (lower is better). The analysis includes 152 of 153 participants whose regularized model fits satisfy the criterion. Effective degrees of freedom are estimated as described in Appendix C.4.2. See Appendix C.5.1 for the Bayesian model shorthands.

C.7.4 Subject-level Held-out Model Comparisons

We compared held-out NLL of FreeE+FreeP with that of the other Bayesian variants and the parametric weighting models at the individual level. For each subject and competing model, we computed the per-trial NLL difference relative to FreeP+FreeE, as defined in Appendix C.5.3. Positive values favor FreeP+FreeE. These comparisons use the same held-out test set (fold 0) as the aggregate model comparisons. We obtained 95% confidence intervals for the median paired differences by resampling participants with replacement 20,000 times.

Table 5: **DMR Choice Task:** Subject-level held-out NLL comparisons across 153 participants. Positive per-trial differences favor FreeP+FreeE. Confidence intervals are obtained from 20,000 participant-level bootstrap samples.

Competing model	FreeP+FreeE wins	Median ΔNLL/trial	95% CI
FreeP+UniformE	128/153 (83.7%)	0.013	[0.010, 0.017]
FreeP+PriorMatchedE	110/153 (71.9%)	0.008	[0.005, 0.010]
FreeP+BoundedLOE	124/153 (81.0%)	0.008	[0.005, 0.012]
LILO	116/153 (75.8%)	0.015	[0.007, 0.018]
Prelec	120/153 (78.4%)	0.014	[0.008, 0.018]

As shown in Table 5, FreeP+FreeE achieves lower held-out NLL for 71.9%–83.7% of participants across the five comparisons. The median per-trial differences are positive in all cases, and all bootstrap confidence intervals do not include zero. These results show that the advantage of FreeP+FreeE is broadly shared across subjects.

C.7.5 Analysis of Fitted Prior and Resources

We show the group-level encoding resources (Figure 27) and priors (Figure 28). Across models, the fitted resources consistently U-shaped, with peaks at 0 and 1. Interestingly, the freely fitted priors for both gains and losses are also U-shaped.

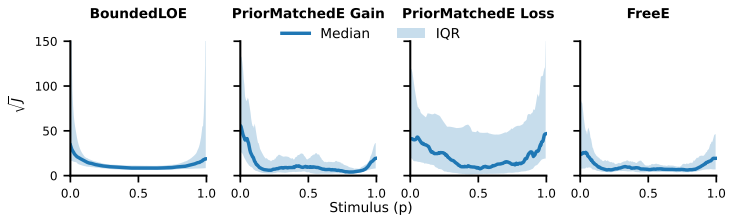


Figure 27: **DMR Choice Task:** Group-level resources ($\sqrt{J(p)}$) for the four encoding components evaluated in this task. The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3.

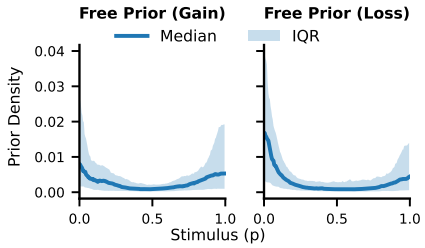


Figure 28: **DMR Choice Task:** Group-level prior of the Bayesian model with freely fitted priors and encoding. The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3.

C.7.6 Analysis of Bias

Figure 29 shows the per-subject bias for the Bayesian freely fitted model, the LILO model, and the Prelec model. Across subjects, the three models tend to capture similar patterns of bias, although the detailed shapes at the individual level likely reflect a considerable degree of overfitting.

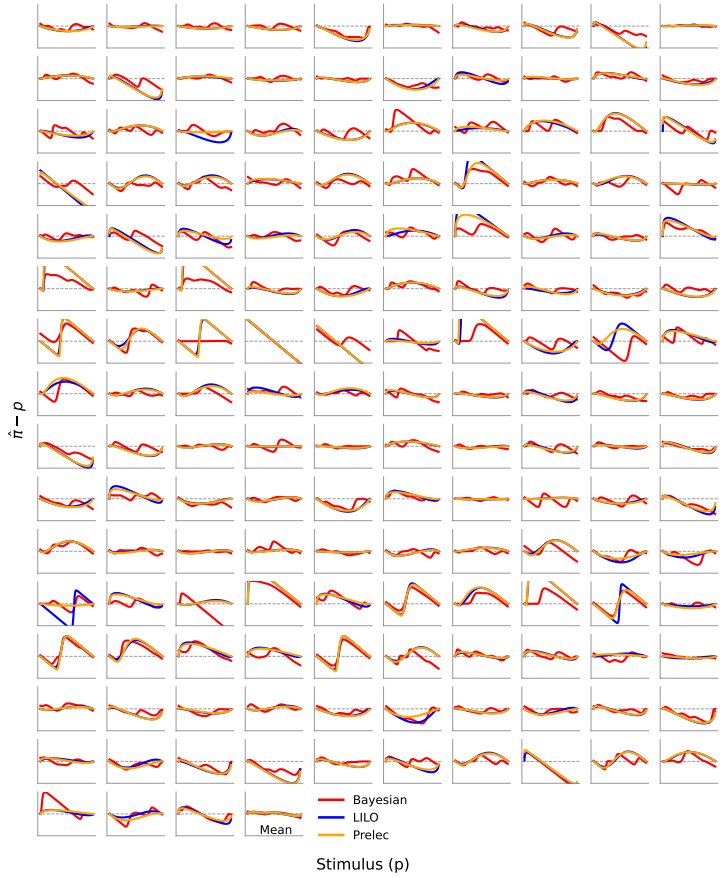


Figure 29: **DMR Choice Task:** Per-subject bias of the LILO model, the Prelec model and the Bayesian model with free prior and free encoding on the DMR pricing task. Subjects S1 to 153 are from dataset CPC15 [6].

C.8 Details on Bayesian Model for Adaptation Bimodal Data

C.8.1 Bayesian Model Variants

We tested several model variants. Most aspects were the same as the variants used for the JRF task, with the exception that here we evaluated a bimodal prior instead of a uniform prior.

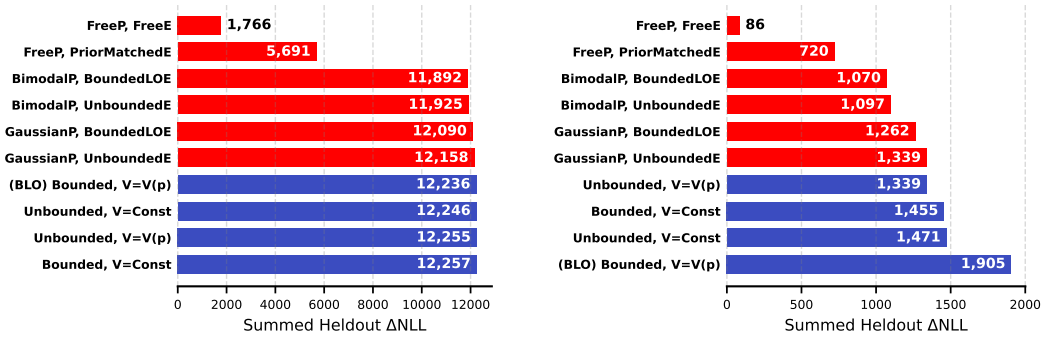
- **Priors:**
  - **Bimodal Prior:** This variant assumes a bimodal distribution over the range of possible grid points. The free parameters are the two means and their corresponding standard deviations.
  - **Gaussian Prior, Freely Fitted Prior:** These variants are the same as described in Appendix C.5.1.
- **Encodings:**
  - **Unbounded Log-odds Encoding, Bounded Log-odds Encoding, Prior-matched Encoding, Freely Fitted Encoding:** Same as in Appendix C.5.1.

C.8.2 Performance Comparison of All Model Variants

Figures 30 and 31 compare model performance in the adaptation task using held-out ΔNLL and effective-df-adjusted ΔAICc under both analytical and categorical likelihood specifications.

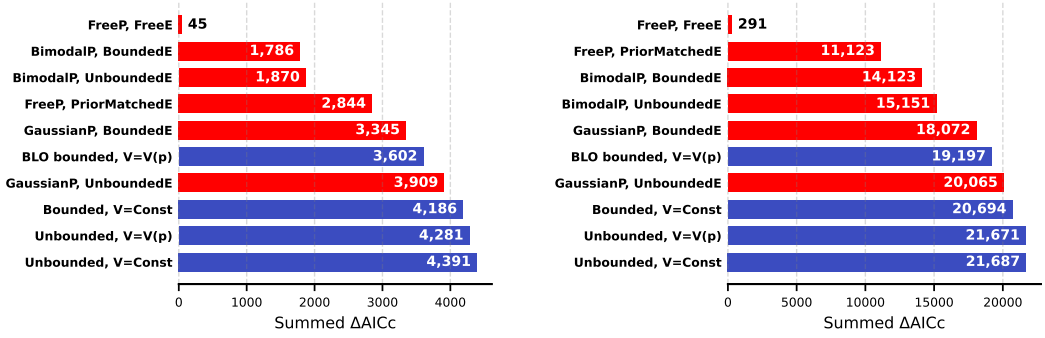
FreeP+FreeE achieves the lowest summed held-out ΔNLL among all models under both specifications.

The complementary effective-df-adjusted AICc comparisons favor FreeP+FreeE under both likelihood specifications as well. These analyses are restricted to participants whose fits satisfy the stationarity criterion (Appendix C.4.2), retaining 9 of 26 participants under the analytical likelihood and 23 of 26 under the categorical likelihood. The median effective degrees of freedom for FreeP+FreeE are 19.7 [IQR: 15.8, 29.3] and 45.8 [IQR: 22.9, 48.7]. FreeP+FreeE achieves the lowest participant-level AICc for 8/9 and 21/23 retained participants. However, the analytical AICc comparison should be interpreted cautiously given its limited sample coverage.



(A) Analytical (B) Categorical

Figure 30: **Adaptation Task:** Model comparison using summed held-out ΔNLL (lower is better). Stimuli follow a bimodal distribution. See Appendix C.5.1 for the shorthands for Bayesian models (red), and Zhang et al. [30] for the corresponding model variants (blue). Analytical and categorical likelihood specifications are described in Appendix C.4.3. The main-text results correspond to the analytical likelihood. Model rankings differ somewhat between the two likelihood specifications; see Appendix C.8.2 for discussion.



(A) Analytical (B) Categorical

Figure 31: **Adaptation Task:** Model comparison using summed effective-df-adjusted ΔAICc (lower is better). Stimuli follow a bimodal distribution. See Appendix C.5.1 for the shorthands for Bayesian models (red), and Zhang et al. [30] for the corresponding model variants (blue). Analytical and categorical likelihood specifications are described in Appendix C.4.3. Comparisons are restricted to subjects whose regularized model fits satisfy the stationarity criterion: 9/26 for the analytical likelihood and 23/26 for the categorical likelihood. Effective degrees of freedom are estimated as described in Appendix C.4.2.

C.8.3 Subject-level Held-out Model Comparisons

We compared held-out NLL across the 26 subjects using the analytical likelihood and the same test split (fold 0) as in the aggregate comparisons. For each participant, we calculated the per-trial NLL difference relative to FreeP+FreeE, as defined in Appendix C.5.3. Positive differences favor FreeP+FreeE. Table 6 reports the number of subjects favoring FreeP+FreeE and the median per-trial differences, with percentile 95% confidence intervals from 20,000 participant-level bootstrap resamples.

Table 6: **Bimodal Adaptation Task:** Subject-level held-out NLL comparisons. Positive per-trial differences favor FreeP+FreeE. Confidence intervals are for the median difference across subjects.

Competing model	FreeP+FreeE wins	Median ΔNLL/trial	95% CI
FreeP+PriorMatchedE	18/26 (69.2%)	0.140	$[-0.023, 4.656]$
BLO	26/26 (100%)	5.286	[1.205, 6.135]

FreeP+FreeE outperforms BLO for all 26 subjects. Against FreeP+PriorMatchedE, it performs better for 18 subjects, although the confidence interval for the median difference includes zero.

We also tested whether FreeP+FreeE's overall advantage in held-out NLL was robust to resampling participants. We resampled the 26 participants with replacement 10,000 times and, for each sample, summed the held-out NLL differences between FreeP+FreeE and each competing model. Table 7 reports the 95% confidence intervals and the proportion of bootstrap samples in which FreeP+FreeE has the lower summed NLL.

Table 7: **Bimodal Adaptation Task:** Bootstrap analysis of summed held-out NLL differences under the analytical likelihood. Positive differences favor FreeP+FreeE. Results are based on 10,000 participant-level bootstrap resamples.

Competing model	Bootstrap support	95% CI of summed ΔNLL
FreeP+PriorMatchedE	98.1%	[233.43, 7550.74]
BLO	100.0%	[7518.55, 13399.81]

The summed NLL difference favors FreeP+FreeE in 98.1% of bootstrap samples against FreeP+PriorMatchedE and in all samples against BLO. Bootstrap support was also 100% against the remaining seven model variants.

C.8.4 Robustness of Model Comparison via Bootstrapping

We performed a subject-level bootstrap analysis on the aggregate model comparison results (Figure 6D). Specifically, we resampled the 26 subjects with replacement 10,000 times. For each bootstrap sample, we calculated the Summed Held-out ΔNLL difference between our freely fitted Bayesian model (FreeP+FreeE) and each competing model.

Table 8 summarizes the results. We report the Bootstrap Support, the percentage of resampled datasets in which the FreeP+FreeE model achieved a lower Summed NLL than the competitor. We also report the 95% Confidence Interval (CI) of the difference in NLL (competing model minus FreeP+FreeE).

The freely fitted Bayesian model significantly outperformed all competing accounts, including the PriorMatchedE and the BLO model, achieving a bootstrap support greater than 98%. The 95% confidence intervals for the NLL differences exclude zero for all comparisons. This analysis confirms the superiority of our Bayesian account in this task.

Table 8: **Adaptation Task:** Participant-level bootstrap comparison of summed held-out NLL across 10,000 resamples. FreeP+FreeE is the reference model; positive ΔNLL values favor FreeP+FreeE. Bootstrap support denotes the percentage of resamples with positive ΔNLL. Confidence intervals are the 95% bootstrap intervals for the summed NLL difference.

Competing model	Bootstrap support	95% CI of ΔNLL
FreeP+PriorMatchedE	98.1%	[233.43, 7550.74]
BimodalP+BoundedLOE	100.0%	[7177.88, 13032.97]
BimodalP+UnboundedE	100.0%	[7216.18, 13054.51]
GaussianP+BoundedLOE	100.0%	[7380.49, 13224.09]
BLO (Bounded, $V = V(p)$)	100.0%	[7518.55, 13399.81]
GaussianP+UnboundedE	100.0%	[7466.34, 13271.41]
BLO (Bounded, $V = \text{Const}$)	100.0%	[7532.97, 13426.86]
BLO (Unbounded, $V = V(p)$)	100.0%	[7518.60, 13437.80]
BLO (Unbounded, $V = \text{Const}$)	100.0%	[7510.38, 13429.20]

C.8.5 Analysis of Fitted Prior and Resources

Figure 32 shows that the group-level resources largely retain the U-shape across encoding variants, consistent with the JRF task results. Figure 32 indicates that the subject-level resources are high near one or both boundaries. The U-shaped encoding holds broadly in the adaptation task. Subject-level resources in Figure 33 also show obvious one-side or two-side elevated values near the boundaries.

Because each subject was exposed to a different bimodal stimulus distribution, we do not plot a group-level prior. Figure 34 shows the freely fitted prior for each subject, which for most subjects successfully adapts to the underlying bimodal distribution.

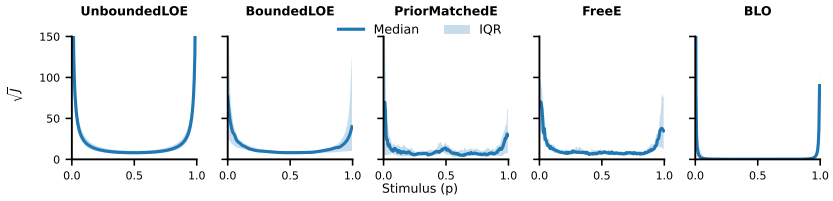


Figure 32: **Adaptation Task:** Group-level resources ($\sqrt{J(p)}$) for the three encoding components evaluated in this task. The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3.

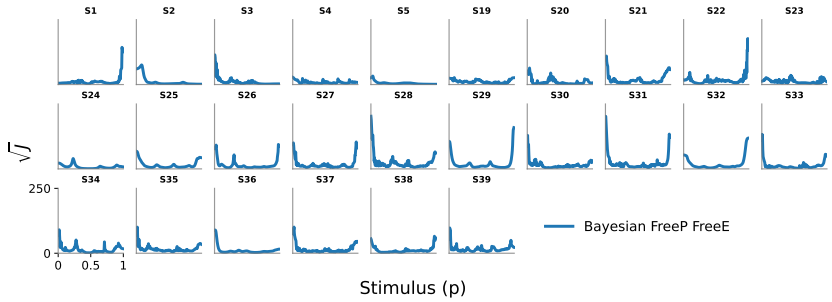


Figure 33: **Adaptation Task:** Fitted resources ($\sqrt{J(p)}$) for individual subjects.

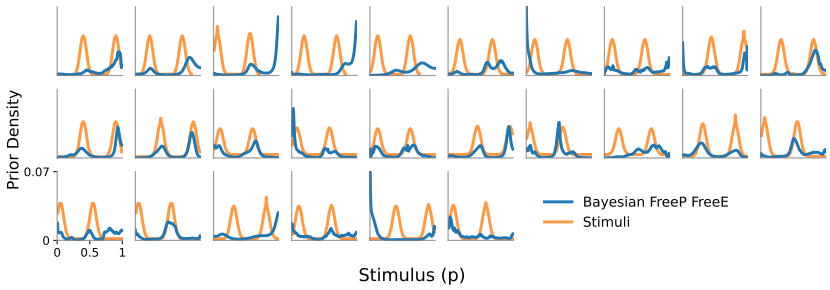


Figure 34: **Adaptation Task:** Fitted priors for individual subjects. The freely fitted prior (blue) successfully adapts to the bimodal stimulus distribution (orange).

C.8.6 Analysis of Bias and Variance

Figure 35 shows the bias of each subject. The BLO model captures broad trends in bias but lacks the flexibility to account for the more complex, stimulus-dependent bias patterns observed in human data. In contrast, the Bayesian model with freely fitted prior and encoding provides a closer fit to subject-level biases, particularly in regions where deviations from linearity are more pronounced.

In the variability figure(Figure 36)The Bayesian model also provides a superior account of variability compared to BLO. While BLO could capture the overall variability magnitude, the Bayesian model models both the overall magnitude and the shape of variability more accurately, and aligns more closely with the human data.

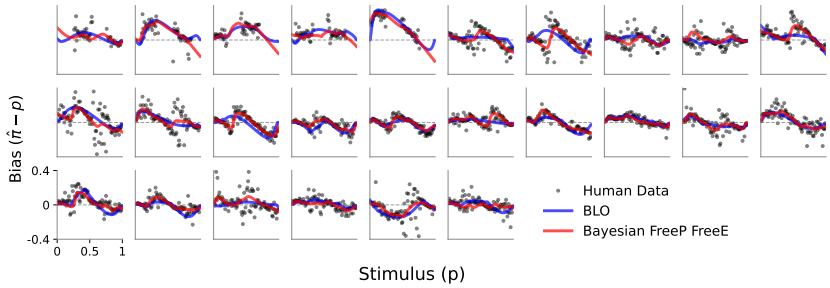


Figure 35: **Adaptation Task:** Per-subject bias of the non-parametric estimates from the data, the BLO model and the Bayesian model with free prior and free encoding on the Adaptation task. Subjects S1 to S26 are from our collected dataset.

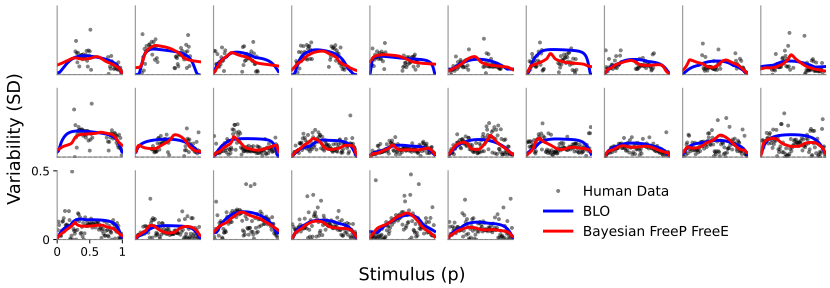


Figure 36: **Adaptation Task:** Per-subject variability of the non-parametric estimates from the data, the BLO model and the Bayesian model with free prior and free encoding on the Adaptation task. Subjects S1 to S26 are from our collected dataset.

C.8.7 Sensitivity to the Decoding Rule

To assess the sensitivity of the recovered prior and encoding to the decoding rule, we refitted the bimodal adaptation data from all 26 subjects using posterior-median, instead of the default posterior-mean decoding.

At the group level, the recovered encoding resources remain U-shaped under both decoding rules (Figure 37). At the subject level, many subjects also exhibit elevated resources near the boundaries (Figure 38). The recovered priors also broadly reflect the bimodal stimulus distribution, with many subjects exhibiting increased prior density near one or both stimulus modes (Figure 39).

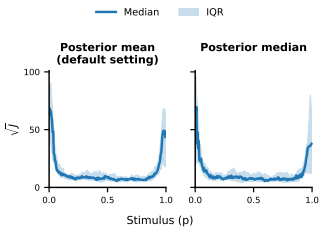


Figure 37: **Bimodal Adaptation Task:** Group-level encoding resources ($\sqrt{\mathcal{J}(p)}$) under posterior-mean and posterior-median decoding across 26 subjects. Solid lines indicate the group median, and shaded regions show the interquartile range (IQR). Both decoding rules yield broadly U-shaped resource profiles.

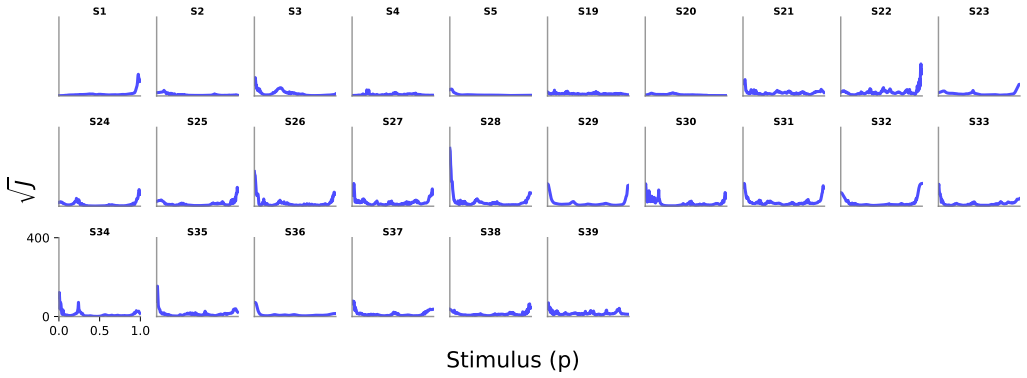


Figure 38: **Bimodal Adaptation Task:** Individual-subject encoding resources ($\sqrt{\mathcal{J}(p)}$) recovered under posterior-median decoding for all 26 subjects. Many profiles exhibit elevated resources near the boundaries, with substantial individual variability.

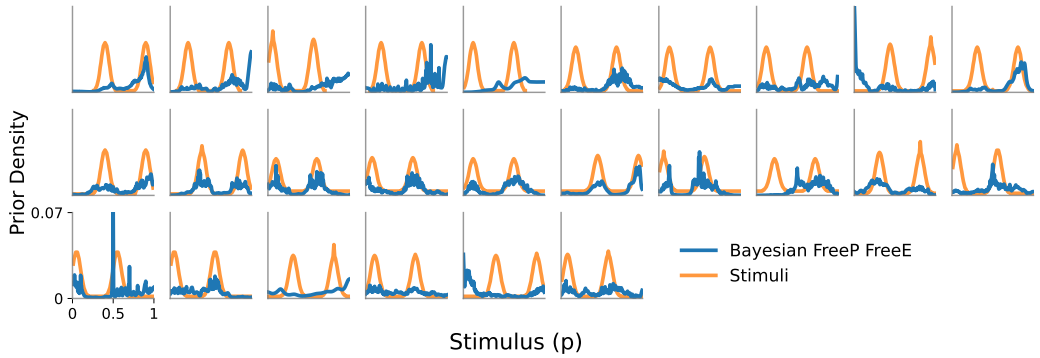


Figure 39: **Bimodal Adaptation Task:** Individual-subject priors recovered under posterior-median decoding for all 26 subjects (blue), compared with the bimodal stimulus distribution (orange). Recovered prior shapes vary across subjects and do not uniformly reproduce the two stimulus modes.

C.9 Cross-Task Analysis

C.9.1 Bias Decomposition

As shown in Section 3.2, we decomposed the total estimation bias ($\mathbb{E}[\hat{p}|p] - p$) into three components: Prior Attraction, Likelihood Repulsion, and Boundary Regression.

We computed these components by the methods describe in Appendix Section C.11.1.

Figure 41 illustrates the bias decomposition for the DMR Pricing task. Similar to JRF, the UniformE model exhibits zero likelihood repulsion. It attempts to compensate for this absence by inducing a heavily distorted Prior Attraction and relying on Boundary Regression. While this compensation allows for a moderate quantitative fit in pricing (a task where we compare the expected value instead of probability), it remains mechanistically distinct from the repulsion in U-shaped models, ann we seek for more evidence in Figure 42, the bias decomposition for the DMR Choice task. As in other tasks, the Uniform model shows zero likelihood repulsion. However, unlike in the Pricing task, the Prior Attraction here remains relatively flat and fails to compensate for the missing repulsion. This flatness likely occurs because, in a binary choice task, increasing prior precision (to reduce variance) without sufficient sensory discriminability at the boundaries would only compress values and reduce the model's ability to predict preference reversals. The model thus defaults to a smoother prior to avoid introducing excessive bias error.

Figure 43 shows how the Bayesian model captures the shift in behavior under bimodal stimulus statistics. In Prior Attraction, the GaussianP model produces a simple, smooth attraction curve. In contrast, both the FreeP and BimodalP models exhibit multi-peaked attraction patterns. These fluctuations correspond to attraction forces directed toward the two modes of the bimodal stimulus distribution.

C.9.2 Resources and Priors

The freely fitted priors and corresponding resources functions in all tasks are shown in Figure 44). Priors differ between tasks, but resources consistently exhibits the U-shape, with highest sensitivity near the extremes. Interestingly, in the DMR Choice tasks, the shape of the prior appears more closely aligned with the resources than in the other tasks. This prior–resources match is reminiscent of the account proposed by Frydman and Jin [10], but the fact that it is not observed across all tasks means that such alignment might be limited in generality.

C.10 Implementation Details of BLO Baselines

C.10.1 JRF and Adapation

While initially developed for decision under risk (DMR) tasks, the BLO model can be adapted to account for perception in judgment of relative frequency (JRF) tasks. In the JRF context, the core

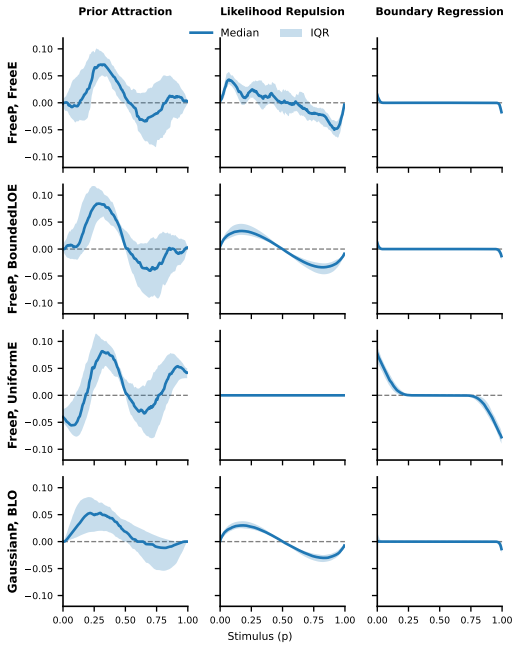


Figure 40: **JRF Task:** The comparison between bias decomposition of four Bayesian model variants. The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3.

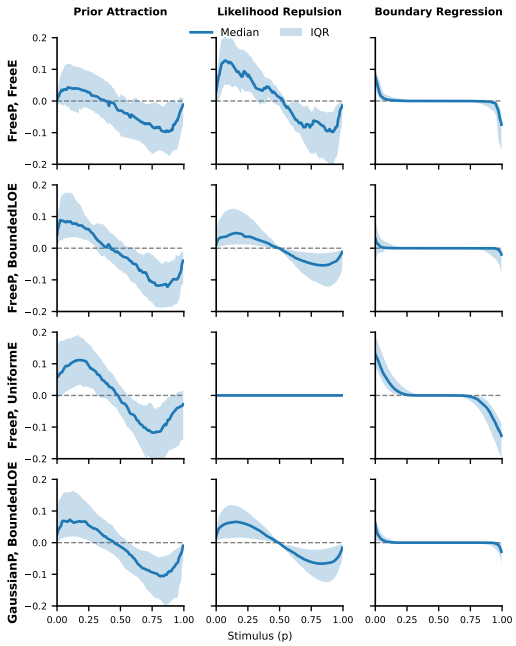


Figure 41: **DMR Pricing Task:** Bias decomposition of four Bayesian model variants. The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3.

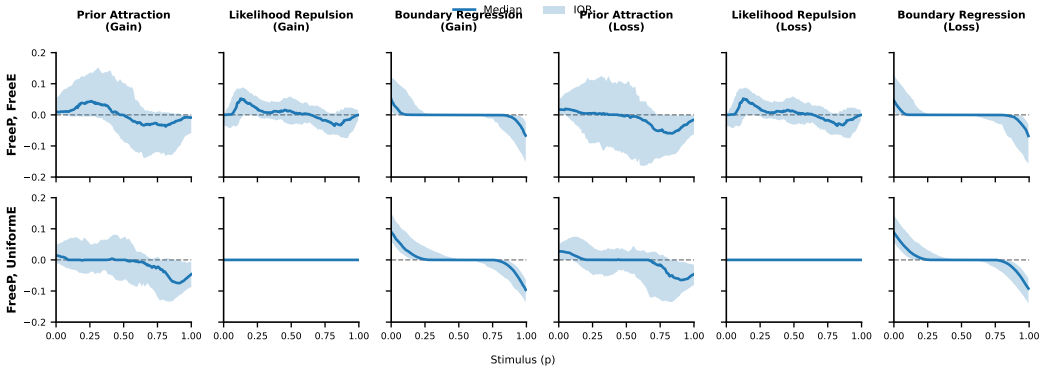


Figure 42: **DMR Choice Task:** Bias decomposition of two Bayesian model variants. The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3.

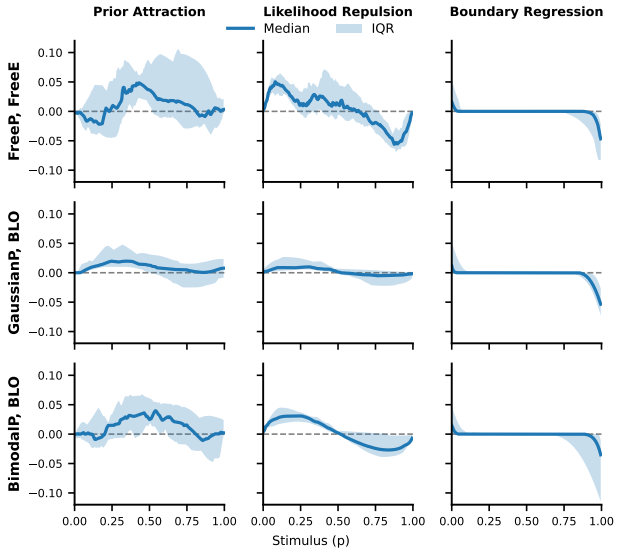


Figure 43: **Adaptation Task:** Bias decomposition of three Bayesian model variants in Figure 6D. The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3.

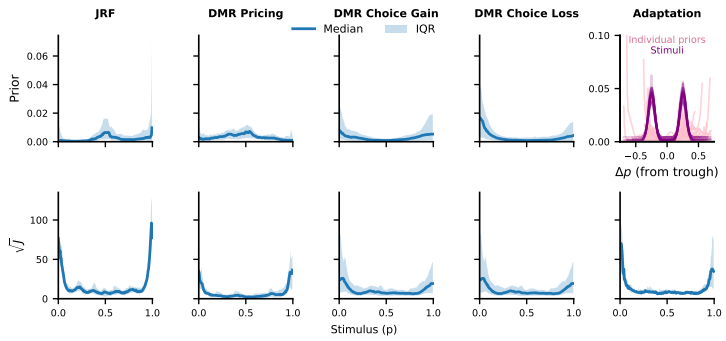


Figure 44: **Across Tasks:** Freely fitted priors (top) and resources $\sqrt{J}$ (bottom) across tasks.The solid line shows the group median, and the shaded area indicates the interquartile range (IQR). Plotting details are provided in Appendix C.11.3. The adaptation prior plotting follows the Figure 6 in the main text.

principles of bounded log-odds encoding, truncation, and variance compensation remain, but the model additionally accounts for sensory noise introduced by the observer's sampling of stimuli.

In JRF tasks, it is assumed that people do not process every element in the display. Instead, they sample a subset of items from the whole display. This sampling introduces additional noise in the observer's estimate. Consequently, the variance of the sample-based estimate, $V(\hat{p})$, is modeled with a finite population correction: If the total number of items is N and the observer samples n_s items, the variance of the sample-based estimate is given by:

$$V(\hat{p}) = \frac{p(1-p)}{n_s} \cdot \frac{N - n_s}{N - 1} \tag{47}$$

, where N is the total number of dots and n_s is a free parameter. We note that sampling of the items is not implemented in the published implementation from Zhang et al. [30], and we correspondingly do not attempt to model it in our reimplementation of the BLO model.

Reimplementation. We reimplemented the key model variants proposed by Zhang et al. [30] for JRF. Specifically, we focused on four variants, Bounded Log-Odds(BLO) and Linear Log-Odds (LLO) models, each combined with either a constant perceptual variance (V=Const) or a proportion-dependent variance compensation (V=V(p)).

Following Zhang et al. [30], we fit these models to the data per subject by minimizing the negative log-likelihood of the observed responses given the model's predicitions. Our reimplementation achieved nearly identical estimated parameter values and model fits to those reported by Zhang et al. [30] for JRF.

For all these reimplemented models, the optimization was performed using a gradient-based approach(Adam or SignSGD) rather than the original Nelder-Mead optimizer(fminsearchbnd) used in their implmentation. We set parameter bounds for our optimization based on Zhang et al. [30]'s settings. the hyperbolic tangent (tanh) function was applied to parameters such as Δ^- and Δ^+ , while the exponential function was used for strictly positive parameters like κ. Parameters not naturally constrained were optimized directly.

We successfully achieved nearly identical estimated parameter values and model fits to those reported by Zhang et al. [30] for the dot counting task, thereby validating our reimplementation.

C.10.2 DMR Pricing

The uncertainty or variance associated with this internal encoding is modeled as being proportional to the binomial variance.

$$V(\hat{p}) \propto p(1-p) \tag{48}$$

Following Zhang et al. [30], the estimate $\hat{\pi}(p)$ is integrated into Cumulative Prospect Theory [25] to predict choice behavior; the certainty equivalent (CE) for a two-outcome lottery $(x_1, p; x_2, 1-p)$ is

given by:

$$\text{CE} = U^{-1}\left[U(x_1)\cdot\hat{\pi}(p) + U(x_2)\cdot(1-\hat{\pi}(p))\right] + \varepsilon_{CE} \tag{49}$$

Here, $U(\cdot) = x^{\alpha}$ is the utility function, and ε_{CE} represents Gaussian noise on the CE scale.

Reimplementation. We first reimplemented the four main parametric models proposed by Zhang et al. [30] for the DMR task. These included the BLO and LLO models, each combined with either a proportion-dependent variance compensation (V(p)) or a constant variance (Const V). As with the JRF task analysis, these models were fitted to the data for each subject individually. Our implementation utilized gradient-based optimizers (specifically, Adam or SignSGD) for parameter optimization.

We observed that the negative log-likelihood (NLL) values obtained from our reimplemented models were nearly identical to Zhang's reported fitted results for most variants. However, for the BLO + V(p) model, our reimplemented model showed, on average, a 7.45 higher loss (negative log-likelihood) compared to Zhang's original results; though this does not impact quantitative model comparison with the Bayesian model.

C.11 Additional Analysis and Plotting Methods

C.11.1 Methods for Generating Empirical Bias Decomposition

Theorem 1 provides an asymptotic bias decomposition for the bounded probability interval $[0, 1]$. To examine the corresponding mechanisms at the fitted noise levels, we use a numerical decomposition based on counterfactual simulations of the fitted Bayesian model. The resulting components sum exactly to the total numerical bias and provide a qualitative counterpart to boundary regression, likelihood repulsion, and prior attraction in the analytical decomposition.

Setup. For each subject and each Bayesian variant, we use the maximum-likelihood parameters $(F, P_{\text{prior}}, \sigma)$ from the cross-validation fit. The forward model produces the bias $\hat{p}(p) - p$ on the 200-point grid over $[0, 1]$.

Counterfactual simulations. To isolate each mechanism, we evaluate the fitted Bayesian model in three configurations:

1. **Full model on $[0, 1]$.** The total bias $b_{\text{total}}(p)$ from the fitted model on the original grid.
2. **Full model on expanded grid.** The grid is widened to $[-1, 2]$ with F' and P_{prior} extrapolated by their endpoint values outside $[0, 1]$. The resulting bias $b_{\text{expanded}}(p)$ reduces the influence of the original boundaries, and the encoding can now spread beyond them.
3. **Uniform-prior model on expanded grid.** Same as (2) but with P_{prior} replaced by a uniform distribution on $[-1, 2]$. The resulting bias $b_{\text{repulsion}}(p)$ contains only the contribution from non-uniform encoding.

Components. The three components are

$$\text{Likelihood Repulsion}(p) = b_{\text{repulsion}}(p), \tag{50}$$

$$\text{Prior Attraction}(p) = b_{\text{expanded}}(p) - b_{\text{repulsion}}(p), \tag{51}$$

$$\text{Boundary Regression}(p) = b_{\text{total}}(p) - b_{\text{expanded}}(p), \tag{52}$$

so that $b_{\text{total}}(p) = \text{Likelihood Repulsion} + \text{Prior Attraction} + \text{Boundary Regression}$.

Mapping to the analytical decomposition. Replacing the prior with a uniform distribution removes the $\frac{1}{\mathcal{J}(p)}\frac{\mathrm{d}}{\mathrm{d}p}\log P_{\text{prior}}(p)$ term in Theorem 1, and expanding the grid reduces the $A_{1,\sigma}(p)\,\text{sign}(0.5 - p)/\sqrt{\mathcal{J}(p)}$ term.

The same procedure is applied to all other tasks and model variants. The resulting bias decomposition figures across tasks and variants are provided in Appendix C.9.1.

C.11.2 Validity of the Small-Noise Approximation

The bias decomposition in Theorem 1 is derived in the small-noise limit. Since the sensory noise levels fitted to human behavior are not necessarily small, we tested whether the analytical approximation remains informative under the fitted parameters.

We compared the analytical approximation with the numerical bias computed directly from the full fitted model. We restricted this comparison to the inner 95% of the probability range, where boundary effects are expected to be weaker. In this region, we used the interior approximation consisting of the likelihood-repulsion and prior-attraction terms:

$$\mathrm{Bias}_{\mathrm{approx}}(p) = \frac{\mathrm{d}}{\mathrm{d}p}\left(\frac{1}{\mathcal{J}(p)}\right) + \frac{1}{\mathcal{J}(p)}\frac{\mathrm{d}}{\mathrm{d}p}\log P_{\mathrm{prior}}(p). \tag{53}$$

We evaluated how well this approximation predicted the sign of the numerical bias across stimulus probabilities, using balanced sign accuracy to account for differences in the frequency of positive and negative bias. Because the approximation requires derivatives of nonparametrically fitted functions, we applied Gaussian local-linear smoothing before differentiation.

Across the 86 JRF subjects, the median balanced sign accuracy was 0.853 (95% CI $[0.818, 0.914]$), using a smoothing bandwidth of 0.1, so the analytical approximation captures the direction of the bias reasonably well. The results were stable across bandwidths from 0.05 to 0.15. We also repeated the comparison with reduced sensory noise while keeping the fitted encoding and prior fixed. As expected from the small-noise expansion, agreement between the analytical approximation and numerical bias improved as noise decreased.

C.11.3 Details for Plotting Group-level Prior and Resources

For each group-level curve presented in the paper, we plot the median across subjects as the solid line and the interquartile range (IQR; 25th–75th percentile) as the shaded area. These quantities are computed pointwise on the stimulus grid, providing a summary of central tendency and between-subject variability.

C.11.4 Methods for Calculating Bias and Variance

This section describes the procedures used to calculate bias and variance for the non-parametric data, the Bayesian model, and the BLO model for the bias and variance figures (for example, Figure 15 and Figure 16).

Non-Parametric Estimation For JRF task, the non-parametric estimate of relative frequency is defined as

$$\hat{\pi}_{NP}(p) = \frac{1}{m}\sum_{t=1}^{m}\hat{\pi}_t(p),$$

where $\hat{\pi}_t(p)$ denotes the subject's estimate on trial t, $t = 1, 2, \ldots, m$.

For DMR Pricing task, probability weights are inferred from certainty-equivalent responses. For each subject, we assign an independent weight $w_t \in (0, 1)$ to each trial and predict the certainty equivalent as

$$\widehat{CE}_t = \left[w_t x_{1t}^{\alpha} + (1 - w_t)x_{2t}^{\alpha}\right]^{1/\alpha},$$

where x_{1t} and x_{2t} are the lottery outcomes and $\alpha > 0$ is a subject-specific power-utility parameter. We jointly estimate the trial-specific weights, α, and a common Gaussian response-error variance by maximum likelihood. The non-parametric probability-weight estimate at each presented probability p is then obtained by averaging the fitted weights across trials presenting that probability.

Bias is the difference between the mean estimate and the true probability:

$$\mathrm{Bias}_{NP}(p) = \hat{\pi}_{NP}(p) - p.$$

The variance is computed as the sample variance of $\hat{\pi}_t(p)$ across trials of the same p.

BLO Model The BLO model produces deterministic predictions $\hat{\pi}_{BLO}(p, N)$ that depend on numerosity N. To obtain a single prediction per probability p, we average across the five numerosity conditions ($N = \{200, 300, 400, 500, 600\}$):

$$\mathbb{E}[\hat{p}]_{\text{BLO}} = \frac{1}{5} \sum_{N \in \{200,300,400,500,600\}} \hat{\pi}_{\text{BLO}}(p, N).$$

Bias is then defined as

$$\text{Bias}_{\text{BLO}}(p) = \mathbb{E}[\hat{p}]_{\text{BLO}} - p.$$

Variance in BLO arises from Gaussian noise in log-odds space, $\epsilon_\lambda \sim \mathcal{N}(0, \sigma_\lambda^2)$. Mapping back into probability space requires a Jacobian transformation:

$$f(p) = \frac{1}{p(1-p)} \cdot \frac{1}{\sqrt{2\pi\sigma_\lambda^2}} \exp\left(-\frac{(\lambda(p) - \hat{\Lambda}_\omega)^2}{2\sigma_\lambda^2} \right).$$

The probability distribution is normalized, and the variance is computed as

$$\text{Var}_{\text{BLO}}(p) = \mathbb{E}[P^2] - \big(\mathbb{E}[P]\big)^2,$$

where expectations are taken with respect to $f(p)$.

Bayesian Model For the Bayesian model, the estimate given a measurement m is $\hat{p}(m)$. The mean estimate for stimulus p is

$$\hat{p}(p) = \sum_m \hat{p}(m)\, P(m \mid p).$$

Bias is defined as

$$\text{Bias}_{\text{Bayesian}}(p) = \hat{p}(p) - p.$$

Variance has two components:

$$\text{Var}_{\text{Bayesian}}(\hat{p}(m) \mid p) = \underbrace{\sum_m P(m \mid p) \left(\hat{p}(m) - \hat{p}(p)\right)^2}_{\text{sensory variance}} + \underbrace{\sigma^2_{\text{motor}}}_{\text{motor variance}}.$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarise the theoretical and empirical results. We took to only include well-supported claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Yes. We discuss limitations in Section 5, including the single-session design of the bimodal adaptation experiment and open questions about timescale separation between encoding and prior adaptation.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: In the main body of our paper, we provide theorems in Section 3. Detailed derivations and proofs are in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: We provide information of the datasets, and experiments in Section 4. Additional details on fitting and plotting are provided in the Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes. Code is openly available (link in main text), with sufficient instructions to reproduce all reported results. Behavioral data used in Sections 4.1 and 4.2 are publicly available from the cited publications. The bimodal adaptation experiment data (Section 4.3) is not released with the code due to informed-consent restrictions. However, fitted model outputs are provided in the supplementary material to enable verification of all the results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes. Model fitting procedure, grid resolution, regularization, and held-out cross-validation are described in Section 4.1; full hyperparameter values, parameter counts for each model variant, and detailed fitting protocols are provided in Appendix C.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes. We report subject-level held-out model comparisons across all four datasets (Appendix Sections C.5.3, C.6.3, C.7.4, and C.8.3). Across-participant variation in recovered encoding resources, priors, and bias is also summarized using medians and interquartile ranges (IQRs) where applicable. Subject-level plots of resources (Figure 10), recovered priors (Figure 34), bias (Figure 15), and response variability (Figure 16) are also provided in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes. Compute resources are described in Appendix C.4: fits run on consumer-grade GPU or CPU, with per-subject fit time ranging from minutes to hours and total compute on the order of GPU-days. Fits are independent across subjects and trivially parallelizable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes. The bimodal adaptation experiment (Section 4.3) was approved by the Ethical Review Board of the Faculty of Mathematics and Computer Science, Saarland University, with informed consent and appropriate compensation provided to subjects. The other behavioral datasets used are publicly available from prior publications. The paper analyzes aggregated behavioral responses without personally identifiable information, and addresses a basic cognitive science question without foreseeable misuse, surveillance, or discrimination implications.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes. The paper studies a basic question in cognitive science with limited direct societal implications. Potential positive impacts include improved understanding of human decision biases for risk communication and behavioral interventions; we do not foresee negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained models, generative models, or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes. The behavioral datasets used in Sections 4.1 and 4.2 are publicly available from the cited publications and used for academic research consistent with their terms of use. The fitting method follows the framework of Hahn and Wei [12], Hahn et al. [13] and is properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes. The bimodal adaptation experiment is documented in Section 4.3 and Appendix C.3. Fitting code with documentation is released alongside the paper, with fitted model outputs in the code base.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to subjects and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification:Yes. Full instructions and compensation details for the bimodal adaptation experiment (Section 4.3) are provided in Appendix C.3.1.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study subjects, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Yes. The bimodal adaptation experiment was approved by the Ethical Review Board of the Faculty of Mathematics and Computer Science, Saarland University (Section 4.3). The study posed minimal risk to participants (a brief visual estimation task with no expected discomfort), and no personally identifiable information was collected. Participants were informed of the voluntary nature of participation and their right to withdraw at any time through the consent form (full text in Appendix C.3.1).

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as a component of the core methods in this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.