THE BOUNDED RATIONALITY OF PROBABILITY DISTORTION: THE BOUNDED LOG-ODDS MODEL

HANG ZHANG[1,2,3] XIANGJUAN REN[3] LAURENCE T. MALONEY[4,5]

1. School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China
2. PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing, China
3. Peking-Tsinghua Center for Life Sciences, Beijing, China

4: Department of Psychology, New York University, New York, NY

5: Center for Neural Science, New York University, New York, NY

Corresponding author:

Hang Zhang
School of Psychological and Cognitive Sciences
Peking University
5 Yiheyuan Road, Haidian
Beijing 100871, China

Email: hang.zhang@pku.edu.cn

Running head: Bounded log-odds model

Counts:

- Abstract:	246 words
- Main text:	12799 words
- Appendices:	1395 words
- Figures	8
- Tables	1
- Supp Figures	7
- Supp Tables	3
- References	122

Author Note

We would like to thank Richard Gonzalez and George Wu for sharing their dataset with us. H.Z. was supported by grants 31571117 and 31871101 from National Natural Science Foundation of China and funding from Peking-Tsinghua Center for Life Sciences. L.T.M. was supported by grant EY019889 from the National Institutes of Health, the Humboldt Research Prize of the Alexander v. Humboldt Foundation, a Fulbright Award and a Guggenheim Fellowship.

Some of the models and results described here were presented as talks at the following conferences or workshops: Psychonomic Society Annual Meeting (November 2013, Toronto, ON, Canada), Workshop on Mathematical and Computational Modeling in Vision (May 2015, St. Pete Beach, FL, USA), Chinese Biophysics Congress (November 2015, Kunming, China), Forty-First Annual Interdisciplinary Conference (February 2016, Breckenridge, CO, USA), Experimental Psychology Meeting of Chinese Psychological Society (May 2016, Chongqing, China), Neuroeconomics workshop at Peking University (June 2017, Beijing, China), "The Probabilistic Brain" workshop at Durham University (March 2018, Durham, UK).

Correspondence concerning this article should be addressed to Hang Zhang, School of Psychological and Cognitive Sciences, Peking University, 5 Yiheyuan Road, Haidian, Beijing 100871, China. Email: hang.zhang@pku.edu.cn

ABSTRACT

In decision under risk, participants act as if their choices were based on probability values systematically different from those that are objectively correct. Similar systematic distortion is found in tasks involving relative frequency judgments and confidence ratings. Observed distortions are highly patterned. They are also *dynamic*: the same participant can have different distortions in different tasks. We propose a model of probability distortion as a form of bounded rationality. This Bounded Log-Odds Model (BLO) is based on the plausible assumptions that (1) the neural representation of probability has a fixed, limited dynamic range ("the bound") and that (2) that objective probability is dynamically remapped to its internal representation to compensate for the limitations imposed by the bound ("rationality"). We test the further hypothesis that (3) the choice of mapping in any task serves to *maximize the mutual information* between objective probabilities in a task and their distorted neural representation. We test these claims in four experiments, two involving judgment of relative frequency, one concerning decision under risk, and a fourth involving both. The last experiment allows us to compare probability distortion in two different kinds of task in a single participant. We demonstrate that BLO accounts for human performance in all of these experiments. We are even able to predict individual choices of probability distortion function across tasks and conditions. Finally, we demonstrate that, subject to these BLO constraints, people optimize their information transmission or nearly so, a form of rational behavior constrained by immutable bounds.

Keywords: probability distortion; frequency judgment; decision under risk; efficient coding; mutual information

In making decisions, we choose among actions whose outcomes are typically uncertain. In the decades surrounding the Second World War a remarkable group of economists and mathematicians (Blackwell & Girshick, 1954; Savage, 1954; von Neumann & Morgenstern, 1944/1953; Wald, 1949) developed methods for modeling decisions as a choice among *lotteries*. To specify a lottery L we list all of its possible outcomes $O_1,...,O_n$ and the corresponding probabilities of occurrence $p_1,\cdots,p_n$ that this specific lottery assigns to each outcome. If we somehow knew all the relevant probabilities, we would be engaged in decision under risk (Knight, 1921).

If we can also assign a numerical measure of utility $U(O_i)$ to each outcome O_i, we could assign an *expected utility* to each lottery,

$$EU(L) = \sum_{i=1}^{n} p_i U(O_i) \ , \tag{1}$$

and a decision maker maximizing expected utility (Bernoulli, 1738/1954; von Neumann & Morgenstern, 1944/1953) would select the lottery with the highest expected utility among those offered: the probabilities serve to weight the contribution of the utility of each outcome. The model is simple but has a wide range of application, not just in economic decisions but also in perception (D. M. Green & Swets, 1966/1974; Maloney & Zhang, 2010) and planning of movement (Trommershäuser, Maloney, & Landy, 2008; Wolpert & Landy, 2012; Wu, Delgado, & Maloney, 2015).

For more than two centuries Expected Utility Theory (EUT) was treated as an adequate description of human choice behavior in decision under risk until challenged by Allais (1953). In an elegant series of experiments he showed that human decision makers did not weight utilities by the corresponding probabilities of occurrence in choosing among lotteries. In Prospect Theory, Kahneman and Tversky (1979) resolved the Allais paradoxes and other shortcomings of EUT by assuming that decision makers use a transformation of

probability $\pi(p)$—a *probability weight* or *decision weight*—in place of probability p in the computation of expected utility.[1] The distortion function in decision under risk $\pi(p)$ was originally inferred from human choices in experiments and it is often but not always an inverted-*S*-shaped function of p (Fig. 1; Gonzalez & Wu, 1999; Tversky & Kahneman, 1992; see Luce, 2000, pp. 84-108 for a comprehensive review of earlier work). The inverted-*S*-shaped form of the function implies that people in effect overestimate small probabilities and underestimate large probabilities.

Wu, Delgado, and Maloney (2009) compared performance in a "classical" decision under risk task with performance in a mathematically identical motor decision task. Each participant completed both tasks and while the fitted probability distortion functions for the classical task were—as expected—inverted-*S*-shaped, those based on the motor task tended to be better fit by *S*-shaped functions. The same participant could have both the inverted-*S*-shaped and *S*-shaped forms of the distortion function $\pi(p)$ in different decision tasks.

Ungemach, Stewart and Chater (2009) examined probability estimates derived from binomial samples and found a similar tendency to underweight small probabilities in decisions and overweight large (see also Barron & Erev, 2003; Fox & Hadar, 2006; Hertwig, Barron, Weber, & Erev, 2004). Probability distortion in the form of inverted-*S*-shaped or *S*-shaped weighting functions is also found in monkeys' choice behavior (Stauffer, Lak, Bossaerts, & Schultz, 2015) and is supported by human neuroimaging evidence (Hsu, Krajbich, Zhao, & Camerer, 2009; Tobler, Christopoulos, O'Doherty, Dolan, & Schultz, 2008).

Zhang and Maloney (2012) reported that both the inverted-*S*-shaped or *S*-shaped distortion functions are found in tasks other than decision-making under risk (DMR). Indeed,

---

[1] The notation $w(p)$ is also used in the literature. Prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) also distinguishes between loss and gain allowing for the possibility that different probability distortion functions are applied in evaluating the impact of potential losses or gains.

people distort probability or relative frequency information in a wide variety of tasks including judgment of relative frequency (JRF), signal detection, confidence rating, etc. (see Fig. 1 for examples). Moreover, Zhang and Maloney found that the same participants had different inverted-*S*-shaped or *S*-shaped probability distortion functions in different experimental conditions even though the trials for the different conditions were randomly interleaved. They concluded that the probability distortion function[2] is not fixed for a participant but *dynamic*, changing systematically with task. Examples from diverse tasks are shown on the top row of Figure 1.

Figure 1 about here

With p and $\pi(p)$ transformed into log-odds, Zhang and Maloney (2012) found that the family of probability distortions could be well fit by a linear transformation:

$$\lambda\left[\pi(p)\right] = \gamma\lambda(p) + (1-\gamma)\lambda(p_0) \ , \tag{2}$$

where $\lambda(p) \ = \log\frac{p}{1-p}$ is the log-odds (Barnard, 1949) or logit function (Berkson, 1944), and $\gamma > 0$ and $0 < p_0 < 1$ are free parameters. The parameter p_0 is the crossover point, the point away from 0 and 1 where the probability distortion function crosses the 45 degree line (Fig. 1abc).

Eq. 1 is called the Linear-in-Log-Odds (LLO) transformation (Gonzalez & Wu, 1999; see Goldstein & Einhorn, 1987; Pitz, 1966; Tversky & Fox, 1995 for earlier uses of LLO). As Figure 1def illustrates, the LLO fit explains a considerable proportion of variance in human data in a diverse set of tasks. See Zhang and Maloney (2012) for further examples, which

[2] For simplicity, we will use the term "probability" to include relative frequency and confidence. The terms “decision weight” or "probability weight" are awkward in tasks where probability information is not used to “weight” utilities or anything else and we avoid them.

include 20 datasets taken from 12 studies involving probability, relative frequency and confidence, all the studies for which we could recover and analyze data.

We caution that the LLO fits to data represent empirical regularities unmotivated by any theory. Over the course of this article we will replace Eq. 2 by a new model based on theoretical considerations. Still, LLO will prove to be an excellent approximation to the new model and any model accounting for the pattern of probability distortion found across tasks should be consistent with LLO. In reporting the results of the experiments, we will report the LLO fits (the observed probability distortion) before testing BLO and its underlying assumptions.

We propose that probability distortion is fundamentally a consequence of a specific limitation on the dynamic range of the neural representation of probability which we identify. As a consequence of this limitation, human performance in a wide variety of tasks (e.g. the Allais Paradoxes) is necessarily sub-optimal by whatever measure is appropriate to the tasks. However, the variation of probability distortion with task hints that—subject to "bounds" on probability representation—individuals might adaptively select the specific form of probability distortion that allows them to perform as well as possible in the task, a form of *bounded rationality* in Herbert Simon's sense (Simon, 1982, pp. 235-244). We will test the Maximum Mutual Information Hypothesis, that the choice of mapping maximizes information between object probabilities and their internal representations subject to the resolution bound on the internal representation of probability.

The remainder of this article is organized into sections as follows:

(1) presentation of the BLO Model,

(2) four experimental tests of the BLO Model and its underlying assumptions in JRF and DMR tasks, including one experiment where the same participants completed both a JRF and a DMR experiment to allow for cross-task comparisons,

(3) comparison of parameters across tasks,

(4) a test of the Maximum Mutual Information Hypothesis,

(5) a discussion of novel predictions of BLO borne out in the empirical results and the limitations of the work presented here.

1. The Bounded Log-Odds Model

There is an imperfect but useful analogy between the Bounded Log-Odds (BLO) Model and lightness adaptation (Gilchrist et al., 1999). The dynamic range of perceived surface lightness from white to black is about 30:1, much less than the dynamic range of light intensities in the terrestrial environment. The visual system dynamically adapts to overall light intensity to maintain some degree of lightness constancy, remapping the input signal to the available dynamic range of the internal representation of lightness.

In effect, the visual system selects a "window" on the terrestrial light field. Reflected light from grey-level surfaces below the window are equivalenced to black, those above are white or self-luminous. Within the window the visual system can discriminate shades of gray. Outside the window discrimination is at best crude. The perceptual system at any specific moment is sensitive only to a limited range of stimuli that it in effect selects.

When you step out of a dark room into sunlight your eyes need seconds or minutes of adaptation before seeing detail in the new illumination setting. The logarithm of physical intensity is mapped onto the internal representation of lightness by a dynamic linear transformation chosen to match the current light environment to the limited internal scale of lightness. There is also increasing evidence that adaptation of representational range occurs along more abstract dimensions, such as value (Kobayashi, Carvalho, & Schultz, 2010; Louie, Khaw, & Glimcher, 2013; Polanía, Woodford, & Ruff, 2019; Rustichini, Conen, Cai, & Padoa-Schioppa, 2017; Tobler, Fiorillo, & Schultz, 2005), numerosity (Burr & Ross, 2008; Cicchini, Anobile, & Burr, 2014), and variance (Payzan-LeNestour, Balleine, Berrada, & Pearson, 2016).

The Bounded Log Odds (BLO) Model has an analogous structure: our internal representation of probability information has a limited range and we adapt our mapping of

probability to the representation to enhance performance. BLO is based on three assumptions which we present next:

1. log-odds representation

2. representation on a bounded Thurstone scale

3. variance compensation

We use factorial model comparison (van den Berg, Awh, & Ma, 2014) to separately test each of the three assumptions against plausible alternatives. We will also compare BLO to previous models found in the literature. The models we considered are not all nested (see below) and factorial model comparison (van den Berg, Awh, & Ma, 2014) does not require nested models.

Assumption 1: log-odds representation.

In the BLO model, we assume probability, p, is internally represented as a linear transformation of log-odds,

$$\lambda(p) = \log \frac{p}{1-p} \ , \tag{3}$$

a one-to-one increasing transformation of probability. Several lines of evidence suggest that log-odds may serve as an internal scale for the neural processing of probability information (see Zhang & Maloney 2012 for review). For example, in a probabilistic inference task, Phillips and Edwards (1966) required participants to respond on a scale of probability, logarithm of probability, or log-odds. They found that participants' deviation from the correct answer was least in log-odds. Maloney and Dal Martello (2006) found that people's similarity ratings for a pair of children faces were a linear function of the log likelihood ratio they judged the children to be siblings, a finding that was later replicated with young adult faces (DeBruine et al., 2009). It is as if people are reading out their kinship judgments in log-odds to use for the facial similarity rating.

Gold and Shadlen (2001, 2002) proposed a neural computation mechanism that allows log-odds to be readily encoded by neuron populations. In BLO, we assume a psychological

scale in log-odds, disturbed by Gaussian random errors. A similar log-odds scale has been introduced by Erev and colleagues (Erev, Wallsten, & Budescu, 1994; Wallsten, Budescu, Erev, & Diederich, 1997) to explain the probability distortion in confidence ratings. Such noise-perturbed psychological scales date back to Thurstone (1927). Log-odds representation itself does not lose information (the log-odds function is invertible) and it provides a scale convenient for linear transformations: doubling a probability of 0.8 would yield 1.6, not a valid probability; doubling a log-odds or adding a constant to it, however, would always lead to a valid log-odds.

Assumption 2: representation on a bounded Thurstone scale

We can think of the bounded Thurstone scale (Thurstone, 1927) as a neural device capable of storing magnitudes within a fixed range, the range -1 and 1. We can encode a magnitude signal s anywhere in this range and later retrieve it. The retrieved value s, however, is perturbed by Gaussian noise with mean 0 and variance σ^2: we might store 0.5 and retrieve 0.63 or 0.48. The schematic Gaussian distributions in Figure 2 capture this representational uncertainty. For simplicity we assume that Gaussian error is independent, identically-distributed across the scale (Thurstone's Case V).

Figure 2 about here

The representational scale is then a device in which we can encode information and retrieve it but now contaminated by Gaussian noise: it models a fallible component of working memory and such Thurstonian models of magnitude representation are commonly used in the study of perception and memory (Allen & Yen, 2001). As a consequence of error, the observer may fail to discriminate encoded signals: the greater the proximity of encoded signals, the greater the chance of error (Oberauer & Lin, 2017).

The units of the scale are arbitrary (we chose -1 to 1 for no good reason) and we can imagine encoding numbers on the scale by a linear or non-linear transformation from another scale such as log-odds. We can pick any interval on the log-odds scale and map it to the Thurstone device. In Figure 2 we illustrate two choices. One represents a small range of the log-odds scale using the full range of the Thurstone device, the other represents a larger range also mapped to the full range of the Thurstone device. The row of Gaussians on the two intervals of the log-odds scale symbolize the encoding-decoding uncertainty induced by the Thurstone scale.

There is an evident tension between the two limitations: the greater the log-odds range that needs to be encoded, the greater the density of the magnitudes along the Thurstone scale, and the greater the chances of confusion of nearby codes and vice versa. If we tried to encode 1000 values in the Thurstone scale in Figure 2 for example with its Gaussian uncertainty as shown, we would be only slightly above chance in judging the order of any two adjacent items. The challenge is to choose a transformation that is most beneficial to the organism.

Our concern is with the representation of probability, specifically in the form of log-odds. The dynamic range of probabilities and log-odds in a decision task can be large. The range of probabilities germane to a given task may also be confined to a small interval within [0,1]. All of the relative frequencies in the Lichtenstein et al. (1978) causes-of-death data set (Zhang & Maloney, 2012, Figure 4) are small (thankfully): only these probabilities need be represented and compared in a task concerning relative frequency of causes of death. Intuitively, in tasks where the “probability gamut” is restricted we could expect a remapping of probability that takes advantage of this restriction. Analogous to adaptive coding in the sensory system, such a remapping can enhance the overall precision of the represented probabilities at essentially no cost. In mathematical notation we select an interval $\left(\Delta^{-},\Delta^{+}\right)$ on

the log-odds scale to be mapped to the full range of the Thurstone scale and in effect we confine the representation of log-odds λ to this interval:

$$\Gamma[\lambda] = \begin{cases} \Delta^-, & \lambda < \Delta^- \\ \lambda, & \Delta^- \le \lambda \le \Delta^+ \\ \Delta^+, & \lambda > \Delta^+ \end{cases}, \tag{4}$$

where, again, Δ^- and Δ^+ denote the lower and upper bounds on the log-odds scale, $\Delta^- < \Delta^+$. We refer to $\Gamma[\lambda]$ as "truncated log-odds". The value Δ^- will be mapped to -1 on the Thurstone scale, the value Δ^+ to 1. The smaller the *half-range* $\Delta = (\Delta^+ - \Delta^-)/2$ the smaller the uncertainty of the encoded and decoded values (Fig. 2) relative to the log-odds scale: if we encode $s \in (\Delta^-, \Delta^+)$ and then retrieve it, it will be perturbed by Gaussian error with standard deviation proportional to Δ.

The remapping functions are like a microscope: the greater the magnification, the more accurately detail can be resolved but at the cost of a smaller field of view. If we know that the probabilities in a task (such as causes of death) are confined to a small interval of the log-odds scale we can use the entire Thurstone scale to represent this interval and effectively reduce the noisiness of log-odd estimates decoded from the Thurstone scale without loss.

Assumption 3: variance compensation

Besides the random error on the representational scale, there could be an additional uncertainty associated with the encoding of probability, depending on the source of the probability. This encoding uncertainty may also vary with the value of probability. Consider, for example, a task where the estimate of probability comes from sampling. Given *n* samples, the variance of the estimated $\hat{p}$ (governed by the binomial distribution) would be (Wackerly, Mendenhall, & Scheaffer, 2008)

$$V(\hat{p}) = \frac{p(1-p)}{n} \ . \tag{5}$$

which varies with p and is almost three times larger for probabilities near 0.5 than those near 0.1. High variance estimates in the center of the probability scale are "less known" than those toward 0 or 1 and, intuitively, deserve less space in the representation.

Models based on Bayesian inference explains numerous biases in human perception and cognition (Petzschner, Glasauer, & Stephan, 2015; Wei & Stocker, 2015). There is considerable experimental evidence that human observers compensate for differences in variance along a perceptual scale in a Bayesian way by a non-linear scaling of values that render the variances roughly uniform. In tasks where decision-makers base their judgments on imperfect estimates of probability, we assume that the encoding process comprises choosing a bounded interval to represent and stabilizing of variance across the intervals.

The resulting variance compensation is a non-linear contraction or expansion around a fixed point Λ_0 on the log-odds scale. For any values of probability p ,

$$\Lambda^{\omega}(p) = \tau\left[\ \omega_p \Lambda(p) + (1-\omega_p)\Lambda_0\ \right] , \tag{6}$$

where $\Lambda(p)$ is, as before, truncated log-odds. The reliability parameter ω_p is inversely related to the variance of the estimate $\hat{p}$ of p in any particular task, which serves to compensate for the variance in estimated probability.[3] The scaling factor τ maps the bounded interval to the fixed Thurstone scale. The form of the equation is analogous to that of the LLO transformation described above and we will explore the relation between LLO and BLO in a later section.

[3] See Oruç, Maloney, and Landy (2003), p. 2466ff for a discussion of optimal cue combination. In the cue combination literature, the terms “precision” and “reliability” refer to the reciprocal of the variance of a cue and we will use them without further comment below.

The second assumption (the Thurstone scale) imposes a limitation on the decision system: probability values encoded as log-odds are perturbed by noise in the encoding and decoding processes. The third assumption (variance compensation) and the choice of interval in Eq. 4 permit the decision system to adapt, to partly compensate for the Thurstonian bound, making best use of the Thurstone scale.

Maximization of mutual information

But one piece is missing. What counts as "rational" in our boundedly rational scheme? What would be the optimal compensation, the choice of probability distortion function for a given task that best compensates for the Thurstone bound? The effect of probability distortion is to lose information about probability. A plausible goal is to preserve as much of this information as possible, to optimize the mutual information (Shannon, 1948a, 1948b; Shannon & Weaver, 1949) between the objective probabilities in a particular task and their noisy internal counterparts.

Within our framework, we are in effect trying to transmit probability information across an error-prone communication channel of bounded capacity (Shannon, 1948a, 1948b; Shannon & Weaver, 1949). There is a considerable body of literature concerning efficient coding in biological perception (Attneave, 1954; Barlow, 1961; Simoncelli & Olshausen, 2001) including compensation for variance (Wei & Stocker, 2015). The key idea is that a more precise coding of one stimulus may come at the cost of a coarser coding of another stimulus. A signal may be favored because it occurs more often or because it is more precisely known. In the last part of this paper, we compare human choice of a probability distortion function to optimal choice maximizing mutual information given the Thurstone bound.

In the next sections we describe two experimental tests of the BLO model involving judgments of relative frequency followed by one experimental test involving DMR. Before doing so we clarify the relation between BLO and LLO. Both will play a role in the analyses.

The relationship between BLO and LLO

We compare LLO (Eq. 2)

$$\lambda\left[\pi(p)\right]=\gamma\lambda(p)+(1-\gamma)\lambda(p_0)$$

to BLO rewritten:

$$\begin{aligned}\Lambda^{\omega}(p)&=\tau\left[\ \omega_p\Lambda(p)+(1-\omega_p)\Lambda_0\ \right]\\&=\tau\omega_p\Lambda(p)+(1-\tau\omega_p)\frac{\tau(1-\omega_p)}{1-\tau\omega_p}\Lambda_0\end{aligned}\quad . \tag{7}$$

Over the range $\left[\ \Delta^-,\Delta^+\right]$ truncated log-odds coincides with log-odds: $\Lambda(p)=\lambda(p)$.

Comparing the equations above, over the range $\left[\ \Delta^-,\Delta^+\right]$ $\tau\omega_p$ replaces γ and $\frac{\tau(1-\omega_p)}{1-\tau\omega_p}\Lambda_0$ replaces $\lambda(p_0)$. In LLO, both γ and p_0 are constants but in BLO, however, the reliability parameter ω_p may vary with the value of p depending on the model of variance appropriate to a given task. If a specific dataset is generated by BLO but fitted by LLO, we would expect that the estimates of γ and p_0 would change with experimental conditions as predicted by BLO. Consequently, we can fit LLO to data and look for the pattern of deviations in the fitted coefficients predicted by BLO, a test of the BLO Model.

2. Testing BLO: Experimental Results

We tested BLO in tasks that differed in the source of probability and whether explicit decision is involved. One task is judgment of relative frequency (JRF), where participants report the relative frequency of one token among an array of two tokens (Varey, Mellers, & Birnbaum, 1990; Zhang & Maloney, 2012). The other task is classic decision-making under risk (DMR), where the probability for each outcome is explicitly stated. Compared with many other tasks (e.g. judging the relative frequency of English letters in written English), both tasks have the advantage that the objective probability can be readily manipulated in the laboratory.

2.1. Experiments 1–2: Judgment of Relative Frequency

Methods

We next describe a judgment of relative frequency (JRF) task and two experiments in which participants judged relative frequency (see Supplemental Fig. S1 for illustrations of the experimental procedures and display settings). Experiment 1 was presented in Zhang & Maloney (2012) and we reanalyze the data here in the BLO framework. Experiment 2 is new.

On each trial of Experiment 1, a display of black and white dots was presented for 1.5 s. Human participants were asked to estimate the relative frequency of black (or white) dots and input their estimate as a number out of 1000. Each participant was randomly assigned to report the relative frequency either for the black or for the white dots. The *relative frequencies* could be 0.01, 0.02, …, 0.99. In each block of size 100, each of the relative frequencies except 0.50 occurred once and 0.50 occurred twice.

The total number of dots in a trial (*numerosity*) could be 200, 300, 400, 500, or 600, each of which was repeated 20 times in each block. The order of presentation was randomized. There were 8 blocks of 100 trials. Eleven naïve participants participated. One

additional participant was excluded because of marked inaccuracy in estimating relative frequency (over 10% of her estimates of relative frequency deviated from the true values by at least 0.4).

The task of Experiment 2 was the same as Experiment 1 except that there were two displays[4], presented one after the other at non-overlapping locations, each for 0.7 s. The number of black (white) dots in a display was a binomial random variable with mean set to one of the *relative frequencies* 0.1, 0.2, 0.3,…, 0.9. The total number of dots in the two displays (*numerosity*) could be 16, 24, 36, or 54. In a single trial the two displays had the same numerosity and expected relative frequency. Each relative frequency × numerosity combination was repeated for 3 times in each block. Half of the participants were asked to estimate the average percentage of dots that were black, the other half were asked to estimate the average percentage of dots that were white. Twelve naïve participants (8 male, aged 18 to 31) participated. The first participant completed 5 blocks of 108 trials each and all other participants completed 6 blocks of 108 trials each. One additional participant was excluded because of marked inaccuracy based on the same criteria as used in Experiment 1. The experiment was approved by the University Committee on Activities Involving Human Subjects at New York University. All participants gave written informed consent in accordance with the Declaration of Helsinki.

Applying BLO to JRF

We need additional assumptions when applying BLO to the JRF experiments. One of the key assumptions of BLO is variance compensation and, to apply BLO, we need to specify a model of the participant's sampling process and the variance of the resulting estimates. First, we assume that humans may not have access to all the tokens presented briefly in a display or in a sequence, due to perceptual and cognitive limits (Miller, 1956; Sperling, 1960). Instead, they

[4] The two-display design was originally intended for investigating the integration of probability information over multiple presentations, which is not relevant to our current topic.

take samples from the population and are thus subject to the randomness associated with sampling. Within BLO, probability distortion arises in part from a compensation for the sampling noise captured in our model by the reliability parameter ω_p.

Denote the total number of dots in a display as N and the relative frequency of black dots as p. Suppose a sample of n_s dots is randomly drawn from the display. We assume that the sampling is without replacement. That is, the same dot will not be drawn twice during one sampling, which is reasonable in our case. As a result, the variance of $\hat{p}$ requires a correction for finite population (Cochran, 1977, see Appendix B for the derivation):

$$V(\hat{p}) = \frac{p(1-p)}{n_s}\frac{N-n_s}{N-1} \tag{8}$$

The finite population correction is intuitive: the larger the sample size relative to the population, the smaller the variance. When $n_s = N$, i.e. when the whole population is included in the sample, we should have $\hat{p} = p$ for each sample and thus $V(\hat{p}) = 0$. At the other extreme, when $n_s = 1$, sampling without replacement is equivalent to sampling with replacement, the familiar $p(1-p)$. The BLO variance correction is a weighted mixture of an estimate based on the sample and an "anchor" Λ_0 which may also be stochastic, with its own variance $V(\Lambda_0)$. The optimal weight for combining the two is

$$\omega_p = \frac{V(\hat{p})^{-1}}{V(\hat{p})^{-1} + V(\Lambda_0)^{-1}} \ . \tag{9}$$

Denote $\kappa = V(\Lambda_0)^{-1}$, the *anchor precision parameter*. This equation can be rewritten as

$$\omega_p = \frac{1}{1+\kappa V(\hat{p})} \ . \tag{10}$$

Finally, we assume that encoded values $\pi(p)$ are perturbed by additive Gaussian error (Thurstone, 1927), updating Eq. 7 to

$$\lambda\left[\pi(p)\right] = \tau\left[\omega_p \Lambda(p) + \left(1-\omega_p\right)\Lambda_0\right] + \varepsilon_\lambda , \tag{11}$$

where ε_λ is Gaussian error on the log-odds scale with mean 0 and variance σ_λ^2.

Numerosity effect: data versus BLO prediction

We first used LLO (Eq. 2) to fit participants' estimated relative frequency as a function of the true relative frequency separately for each participant and each numerosity condition. These fits serve as measures that are directly comparable to other reports of probability distortion in the literature (see, for example, those in Zhang & Maloney (2012)). We will see below how probability distortion summarized by these measures is predicted by BLO.

In both experiments, we found that the estimated slope $\hat{\gamma}$ declined as numerosity increased. We call this decline the *numerosity effect*.

Figure 3 about here

The estimated $\hat{\gamma}$ is plotted against numerosity in Figure 3. Each thin black line denotes one individual participant. The thick black line denotes the mean across participants. In Experiment 1[5], the $\hat{\gamma}$ for displays of 600 dots (0.71) was 18% smaller than that of 200 dots (0.87). According to repeated-measures ANOVA, the effect of numerosity on $\hat{\gamma}$ was significant, $F(4, 40) = 11.88$, $P < 0.0001$, $\eta_p^2 = 0.54$. *Post hoc* analyses using Tukey's honestly significant difference criterion at the 0.05 significance level indicated a significant decline in $\hat{\gamma}$ as numerosity increased from 200 to 600.

[5] The statistics reported here for Experiment 1 was slightly different from those of Zhang & Maloney (2012) due to different trial-exclusion procedures: here we excluded no trials.

In Experiment 2, the value of $\hat{\gamma}$ for displays of 54 dots (0.88) was 22% smaller than that for 16 dots (1.13). According to repeated-measures ANOVA, the effect of numerosity on $\hat{\gamma}$ was significant, $F(3, 33) = 16.95$, $P < 0.0001$, $\eta_p^2 = 0.61$. *Post hoc* analyses using Tukey's honestly significant difference criterion at the 0.05 significance level indicated that the reductions in $\hat{\gamma}$ from numerosity 16 to all the larger numerosities were significant, as was the decline from numerosity 24 to numerosity 54.

For both experiments, there were considerable individual differences but all participants had similar declining trends in $\hat{\gamma}$ when numerosity increased (Fig. 3). This trend holds not only for participants who had all $\hat{\gamma}$ less than 1, but also for participants who had some $\hat{\gamma}$ greater than 1, thus ruling out the alternative hypothesis that $\hat{\gamma}$ had a greater deviation from 1 with increasing numerosity.

For the estimated crossover point $\hat{p}_0$, the effect of numerosity was subtle. In Experiment 1, the mean $\hat{p}_0$ for the numerosity of 200, 300, 400, 500, 600 was, respectively, 0.48, 0.55, 0.52, 0.70, 0.70. According to a repeated-measures ANOVA, the difference was only marginally significant, $F(4, 40) = 2.59$, $P = 0.051$, $\eta_p^2 = 0.21$. In Experiment 2, the mean $\hat{p}_0$ for the numerosity of 16, 24, 36, or 54 was, respectively, 0.52, 0.55, 0.61, 0.67. Again, there was an increasing trend but it failed to reach significance, $F(3, 33) = 0.86$, $P = 0.47$, $\eta_p^2 = 0.073$.

We would like to highlight three aspects of the results. First, participants had pronounced probability distortions in this simple relative-frequency judgment task. Second, the slope of distortion can be greater than one as well as less than one. Third, for the same participant, the distortion varied systematically with task condition (i.e. numerosity) even

though trials for all conditions were interleaved. Zhang and Maloney (2012) showed that no existing theory or model could explain all of these results.

For each participant of Experiments 1 and 2, we fit BLO to the participant's estimated relative frequencies using maximum likelihood estimates (see Appendix C for details). We then used the fitted BLO model to generate responses for the experimental stimuli and measured the $\hat{\gamma}$ and $\hat{p}_0$ for this simulated dataset. For each participant, we repeated the simulation 1000 times and computed the mean and 95% confidence interval of $\hat{\gamma}$ and $\hat{p}_0$ for each numerosity condition. For both experiments, for each participant separately, and for both the $\hat{\gamma}$ and $\hat{p}_0$, the numerosity effects predicted by BLO (gray lines and shadings in Fig. 3) closely resembled those of data (black lines, see also Supplemental Fig. S2 for individual fits).

Here is some intuition as to how BLO works. Eqs. 8 & 10 imply that the larger the numerosity N, the smaller the weight ω_p for the original estimate. Since $\gamma = \tau\,\omega_p$ (Eq. 7), this would account qualitatively for the decline of the slope of probability distortion $\hat{\gamma}$ when numerosity increases.

The noticeable (though statistically insignificant) increasing trend with increasing numerosity in the estimated crossover point $\hat{p}_0$ may be explained as an emergent property of Eq. 7: $\lambda(p_0) = \frac{\tau(1-\omega_p)}{1-\tau\omega_p}\Lambda_0$. Since ω_p varies with the numerosity N (Eqs. 8 & 10), $\lambda(p_0)$ would vary even when Λ_0 is unaltered.

Model comparisons

BLO is based on three assumptions. We used *factorial model comparison* (van den Berg et al., 2014) to separately test each of the assumptions, comparing alternative models that differ in the following three "dimensions".

D1: scale of transformation. We considered two alternatives to the log-odds scale: the "Prelec scale" and the linear scale.

The Prelec scale is derived from Prelec's function (Prelec, 1998):

$$\pi(p) = \exp\left(-\beta(-\log p)^{\eta}\right) \tag{12}$$

with η and β as free parameters. LLO and the Prelec families both are among the probability weighting functions that typically fit best to data (Cavagnaro, Pitt, Gonzalez, & Myung, 2013; Stott, 2006). They are difficult to distinguish empirically (Cavagnaro, Pitt, Gonzalez, & Myung, 2013; Luce, 2000).

Taking logarithms and negating twice on both sides of Prelec's function, we can see that Prelec's function is equivalent to a linear transformation

$$\lambda'\left[\pi(p)\right] = \eta\lambda'[p] - \log\beta \tag{13}$$

on the Prelec scale

$$\lambda'(p) = -\log(-\log p)\ . \tag{14}$$

The functional form of the linear scale is based on the *neo-additive family* (Bell, 1985; Birnbaum & Stegner, 1981; Chateauneuf, Eichberger, & Grant, 2007; Cohen, 1992; Gilboa, 1988; Loomes, Moffatt, & Sugden, 2002; Teitelbaum, 2007; see Wakker, 2010 for a review) which refers to a linear transformation of probability except that it may have discontinuities at the extremes to ensure $\pi(p)$ is within [0,1]:

$$\pi(p) = \begin{cases} 0, & if\ p = 0\ or\ kp + b < 0 \\ 1, & if\ p = 1\ or\ kp + b > 1 \\ kp + b, & otherwise \end{cases}, \tag{15}$$

where k and b are free parameters. The linear scale is accordingly defined as

$$\lambda^{\circ}(x) = \begin{cases} 0, & if\ x < 0 \\ 1, & if\ x > 1 \\ x, & otherwise \end{cases} . \tag{16}$$

For models that use the Prelec scale, we simply replaced the log-odds and its reverse transformation with $\lambda'(p)$ and its reverse. For models that use the linear scale, the log-odds transformation was replaced by $\lambda^{\circ}(p)$ but there was no inverse transformation, because $\lambda^{\circ}(p)$ is not invertible, and no need for an inverse transformation, because $\lambda^{\circ}(p)$ corresponds to probability. The noise term ε_{λ} was always added to the log-odds of $\pi(p)$ to maintain a fair comparison between models. D1 does not influence the number of free parameters of the model.

D2: bounded or bounds-free. We assumed a bounding operation in BLO (Eq. 4), where probabilities outside specific boundaries are truncated to the boundaries. We considered alternative models that are bounds-free. Bounds-free models would not include the bounds parameters, Δ^{-} and Δ^{+}.

D3: variance compensation. We considered three different forms for the reliability parameter ω_p in the encoding process (Eq. 6), with the constraint that $0 \le \omega_p \le 1$. In BLO, ω_p is inversely proportional to $V(\hat{p})$ (Eq. 10) so that the encoding variance is appropriately compensated in the framework of Bayesian inference. Given that the $\hat{\gamma}$ measured by LLO is proportional to ω_p (Eq. 7), an alternative way to explain the numerosity effect in $\hat{\gamma}$ is to simply assume that ω_p is a decreasing function of N:

$$\omega_p = \frac{1}{1+\kappa N} \tag{17}$$

for some choice of $\kappa > 0$, as if the compensated variance has the form $V \propto N$. The third form of ω_p we considered is a constant that does not change with p or N at all, as if the compensated variance is constant. When ω_p is constant, the potentially non-linear transformation of BLO (Eq. 7) is reduced to the linear transformation of LLO, as a re-parameterization would reveal (with one free parameter reduced). The three forms of variance compensation will be referred to as $V(p)$, $V \propto N$, and $V = const$.

The three dimensions—D1, D2, and D3—correspond to the three assumptions of BLO.

We did not list the presence or absence of the scaling factor τ as a possible dimension for both theoretical and practical reasons. On one hand, τ is required to map the bounded interval to the fixed Thurstone scale. On the other hand, the absence of scaling would preclude a greater-than-one $\hat{\gamma}$. There is evidence from several laboratories other than our own experiments that $\hat{\gamma}$ can be greater than 1 (Brooke & MacRae, 1977; Pitz, 1966; Shuford, 1961; Wu et al., 2009).

The three dimensions are independent of each other, analogous to the different factors manipulated in an experiment with a factorial design. In total, we tested 3 (D1: log-odds, Prelec, linear) × 2 (D2: bounded, bounds-free) × 3 (D3: $V(p)$, $V \propto N$, $V = const$) = 18 different models. LLO is among the 18 models (when D1=log-odds, D2=bounds-free, D3= $V = const$).

For each participant of Experiments 1 and 2, we fit each of the 18 models to the participant's estimated relative frequencies using maximum likelihood estimation (see Appendix C for details). The Akaike information criterion with a correction for sample sizes (AICc) (Akaike, 1974; Hurvich & Tsai, 1989) was used for model selection. For each participant, we identified the model with the minimum AICc. Δ AICc was defined as the

difference of AICc between the model in question and the minimum AICc. The higher the value of ΔAICc, the worse the model fit to the participant's data.

Figure 4ab shows ΔAICc summed across participants for the 18 models, arranged in the three dimensions. The manipulation of each dimension had visible impact on the summed ΔAICc, in most cases of which the assumptions of BLO had an advantage over alternative assumptions. For both experiments, BLO was the model of the lowest summed ΔAICc.

Figure 4 about here

We would like to see how each of BLO's assumptions behaves when compared to the alternative assumptions on each dimension. To this end, we divided the 18 models into three families by their assumptions on D1—log-odds, Prelec, and linear. The exceedance probability was computed for each model using the group-level Bayesian model selection (Stephan, Penny, Daunizeau, Moran, & Friston, 2009), which indicated the probability for the model to be the best model among the 18 models. Summed exceedance probability of a specific family was defined to be the sum of exceedance probabilities for all the models in the family. We did the same thing for D2 and D3. According to the summed exceedance probabilities (Table 1), the assumptions of BLO outperformed the alternatives to these assumptions on each of three dimensions.

Predicted probability distortions

BLO had a great advantage over LLO in fitting participants' judgment of relative frequency (Fig. 4ab). We visualized how the probability distortion predicted by BLO may differ from that of LLO (Supplemental Fig. S3ab). For LLO, $\pi(p)$ appears as an inverted-*S*-shaped or *S*-shaped function of p. The shape of $\pi(p)$ predicted by BLO differs from that of LLO mainly in the

"elbows" near 0 and 1, the combined result of boundedness and variance compensation. One such elbow is marked with a red arrow in the first panel.

2.2. Experiment 3: Decision under risk

We further tested whether BLO could be applied to modeling probability distortion in decision under risk (DMR). The dataset we considered was from Gonzalez and Wu (1999), where they tested each of 10 participants on 165 different lotteries (15 value sets × 11 probabilities, see their Table 1) and measured the *certainty equivalent*—the value of sure reward that is equally preferred—to each lottery. We will refer to Gonzalez and Wu's (1999) experiment as Experiment 3. In this seminal work, Gonzalez and Wu estimated the utility function and probability weighting function (probability distortion) for each participant using a non-parametric method (i.e. without assuming any specific functional forms for probability distortion or utility).

We chose to fit Gonzalez and Wu's (1999) dataset because the set of lotteries it included is large enough and rich enough to allow for reliable modeling on the individual level—as Gonzalez and Wu demonstrated. According to their non-parametric estimate, the probability weighting functions of most participants were inverted-*S*-shaped. There were marked individual differences. Zhang and Maloney (2012) found that these weighting functions—similar to probability distortion in many other tasks—could be well fit by LLO (median R^2 was 0.94). We found that there were large individual differences in both the estimated slope $\hat{\gamma}$ and crossover point $\hat{p}_0$ in agreement with Gonzalez and Wu.

Applying BLO to DMR

To model $\pi(p)$, BLO's assumptions for different tasks are the same, except that encoding variance is task-specific. The variance to be compensated for JRF was defined in

terms of the binomial variation to be expected in sampling without replacement (Eq. 8). Probability is described explicitly in DMR and there need be no uncertainty about its value. Participants' choices suggested, however, that they were still compensating for some kind of encoding uncertainty that varies with the value of probability. Gaussian encoding noise on the log-odds scale, when transformed back to the probability scale, results in variance that is approximately proportional to $p(1-p)$ (see Appendix A for proof). The reliability parameter in Eq. 6 is thus:

$$\omega_p = \left[1+\kappa p(1-p)\right]^{-1}, \tag{18}$$

where κ is a free parameter. This same equation can be reached if, alternatively, we assume that participants were compensating for a virtual sampling process (the $\frac{1}{n_s}\frac{N-n_s}{N-1}$ term in Eq. 8 can be assimilated into κ for constant N and n_s). Compensation for virtual sampling was also assumed in some previous theories on probability distortion (Fennell & Baddeley, 2012; Martins, 2006). We return to this issue in the Discussion.

Any lottery in Experiment 3 can be written as $(x_1, p; x_2, 1-p)$, where $x_1 > x_2 \geq 0$.[6] For each participant, we modeled the certainty equivalent (CE) of each lottery using Cumulative Prospect Theory (Tversky & Kahneman, 1992) and assumed a Gaussian error term on the CE scale, as did Gonzalez and Wu (1999):

$$CE = U^{-1}\left[\ \pi(p)U(x_1) + \left(1-\pi(p)\right)U(x_2)\ \right] + \varepsilon_{CE}, \tag{19}$$

where $U(.)$ denotes the utility function, $U^{-1}(.)$ denotes the inverse of $U(.)$, $\pi(p)$ denotes the probability distortion function, and ε_{CE} is a Gaussian random variable with mean 0 and

[6] In words, the lottery (x_1,p;x_2,1-p) offers the value x_1 with probability p and otherwise x_2.

variance σ_{CE}^2. The utility function for non-negative gains alone (none of the lotteries involved losses) was assumed to be a power function with parameter $\alpha > 0$:

$$U(x) = x^{\alpha}. \tag{20}$$

Both Cumulative Prospect Theory and the power utility function have been widely used to model decision under risk or uncertainty (Luce, 2000; Wakker, 2010). What of interest here is the functional form (family) of the probability weighting (probability distortion) function, for which LLO and Prelec's function are traditional options[7]. We addressed the same questions as we did in the previous section for the relative-frequency judgment task: Is BLO superior to alternative forms of probability distortion in goodness-of-fit? Do the assumptions of BLO perform better than alternative assumptions on the same dimension?

Model comparisons

The same series of alternative models to BLO could be constructed on the three dimensions—D1, D2, and D3—except that on D3 there would no assumption corresponding to the numerosity N. In total, 3 (D1: log-odds, Prelec, linear) × 2 (D2: bounded, bounds-free) × 2 (D3: $V(p)$, $V = const$) = 12 different models served as the alternative forms of $\pi(p)$ in Eq. 19. We fit the 12 models to the estimated CE in Experiment 3 separately for each participant.

To facilitate the comparison between JRF and DMR, we presented the main results of Experiment 3 in the same format as those of Experiments 1 and 2. Figure 4c shows the summed ΔAICc across participants for the 12 models of Experiment 3. The "bounded Prelec" model with compensation for encoding variance (D1=Prelec, D2=bounded, D3= $V(p)$, i.e. the

[7] For example, Gonzalez and Wu (1999) used LLO as the probability weighting function for their parametric estimation procedure. Using their dataset (Experiment 3), our parameter estimates for the LLO model were the same as those presented in their Table 3.

same as BLO except the replacement of the log-odds scale with the Prelec scale) turned out to be the model of the lowest summed ΔAICc, though its advantage over BLO was small.

The summed exceedance probabilities for each model family in Experiment 3 are shown in Table 1. In accordance with Experiments 1 and 2, the family with the highest summed exceedance probability for D2 and D3 was respectively the bounded and $V(p)$ family. The difference was in D1: the best-fit family in Experiments 1 and 2 was the log-odds family, whereas in Experiment 3 the Prelec family was slightly ahead of the log-odds family.

All the 12 models constructed above for DMR are in the framework of Cumulative Prospect Theory (CPT). We also considered the original Prospect Theory (OPT, Kahneman & Tversky, 1979), where the probability weight for the lower outcome in the gamble, $\pi(1-p)$, is computed in the same way as $\pi(p)$, while $\pi(1-p)=1-\pi(p)$ is assumed in CPT. A parallel set of 12 models were constructed. The results of model comparison among the OPT models (Supplemental Fig. S4a) were similar to those of the CPT models (Fig. 4c), with each OPT model fitting considerably worse than its CPT counterpart.

In addition, we considered a CPT model with Tversky and Kahneman's (1992) weighting function and two competitive models outside the framework of Prospect Theory—the transfer of attention exchange model (Birnbaum, 2005) and Haruvy's regression model (Erev et al., 2010). The transfer of attention exchange model, according to Birnbaum (2005), could outperform CPT in predicting human choices in a wide range of gambles, while Haruvy's regression model was the second best model for the decision from description dataset in Erev et al.'s (2010) model competition (the best model was a variant of CPT). However, for the DMR dataset we tested, the fits of all these models (see Appendix D for model definitions) were either worse than or close to the LLO model and were much worse than the BLO model (Supplemental Fig. S4b).

Predicted failures of stochastic dominance

Supplemental Figure S3c shows the estimated $\pi(p)$ predicted by BLO for each participant in Experiment 3. As in Experiments 1 and 2, bounding gave rise to "elbows" near 0 and 1, though the individual differences of $\pi(p)$ seemed to be much larger than their counterparts in the former two experiments. The large inter-subject variation evident in the fits for Experiment 3 is typical of decision under risk (see Gonzalez & Wu, 1999). This may reflect uncertainty in participants' interpretation of explicit probability statements. Indeed, the following observation suggests that at least one participant had a decidedly non-normative approach to using probability.

We noticed that the BLO estimated $\pi(p)$ of one participant (Supplemental Fig. S3c, thickened light blue curve, Participant 4 in Gonzalez & Wu, 1999) was non-monotonic instead of monotonically increasing: it rose to a local maximum at $p = 0.16$, and then decreased, reaching a local minimum at $p = 0.5$, and then went up again. A non-monotonic $\pi(p)$ may lead to violations of stochastic dominance, i.e. the participant is likely to prefer a lower than a higher probability to receive a reward (Birnbaum, 1992; Mellers, Weiss, & Birnbaum, 1992; Tymula, Belmaker, Ruderman, Glimcher, & Levy, 2013).

At first sight this sounds like a "bug" in the assumptions of BLO—the interaction of bounding and anchoring operations can, in theory, result in a non-monotonic $\pi(p)$ (the $\pi(p)$ predicted by the bounded Prelec model had similar non-monotonicity). To our surprise, a closer examination of Participant 4's estimated CE did provide evidence for violation of stochastic dominance: For two lotteries $(x_1, p; x_2, 1-p)$, where $x_1 > x_2 \geq 0$ are identical, stochastic dominance predicts that the CE for a smaller value of p should be smaller than the CE for a larger value of p. Among the probabilities used in the experiment, the value p = 0.25 is close to the local maximum and the values p = 0.40, 0.50, or 0.60 are close to the local

minimum of the non-monotonic probability distortion curve of the anomalous participant. As the consequence, $\pi(0.25)$ is higher than $\pi(0.40)$, $\pi(0.50)$, or $\pi(0.60)$, and we may expect violations of stochastic dominance, if any, to occur in the comparisons of the CE of p = 0.25 to the CE of p = 0.40, 0.50, or 0.60. For Participant 4, among the 45 pairs of CE comparisons, the CE of p = 0.25 was *larger* in 29 pairs (a match counts as 0.5 pair). According to Fisher's exact test, we could reject the null hypothesis that the CE of p = 0.25 is larger in no more than half of the pairs at a significance level of 0.036. That is, Participant 4's responses significantly violated stochastic dominance as predicted by the BLO model.

2.3. Experiment 4: different tasks, same individuals

In Experiments 1–3, we found that probability distortions in both JRF and DMR can be well modeled by BLO, which suggests a common representation of probability underlying different tasks. However, it is still possible that the probability distortions in different tasks tax distinct mechanisms and their similarities are coincident. As a more direct test of the common representation, we conducted a fourth experiment where each participant completed both the JRF and DMR tasks.

Methods

In Experiment 4, each participant performed two tasks—judgment of relative frequency and decision under risk, whose procedures and designs respectively followed those of Experiments 1 and 3 (payoffs in the gambles were in RMB instead of USD) except for the changes below. In JRF, participants reported their estimate of relative-frequency by clicking on a horizontal bar with tick marks from 0 to 100%, which was more time-saving compared to inputting the percentage using keyboard. To make the two tasks comparable, the objective relative frequency of JRF was chosen from 11 possible values (0.01, 0.05, 0.10, 0.25, 0.4, 0.5,

0.6, 0.75, 0.90, 0.95, 0.99) as its counterpart in DMR. In each session there were 11 (probability) × 15 (outcome pair) = 165 DMR trials and 11 (probability) × 5 (numerosity) × 6 = 330 JRF trials, which took approximately two hours.

Experiment 4 consisted of two sub-experiments, 4A and 4B. In Experiment 4A, trials from the two tasks were randomly interleaved. Participants completed two sessions on two different days, so that we could evaluate the consistency of their performance. For the first 29 participants (13 male, aged 18 to 29) of Experiment 4A, the dots in each display were distributed within a circular area of 12° diameter, which was slightly different from the setting of Experiment 1 (17°×17° square area). When we failed to replicate the numerosity effect of JRF in these participants, we switched to the original visual setting of Experiment 1 and ran 22 more participants (7 male, aged 18 to 25). It turned out the numerosity effect was still absent (see Supplemental Fig. S6a for probability distortion under each numerosity). Then we conjectured that the interleaved DMR trials might cause interference on JFR. This conjecture was verified by Experiment 4B (Supplemental Fig. S6a), where participants completed one task before the other in one session. Experiment 4B had 24 participants (10 male, aged 18 to 27), half of whom did JRF first and the other half, DMR first. Six additional participants from the first run of Experiment 4A were excluded from further data analysis for missing the second session: for four of them, the second session was cancelled by the experimenter based on a performance screening procedure we later found unnecessary; for the remaining two, the second session was cancelled by the participant. No participants were excluded from the second run of Experiment 4A or Experiment 4B. The experiment was approved by the Institutional Review Board of School of Psychological and Cognitive Sciences at Peking University. All participants gave written informed consent in accordance with the Declaration of Helsinki.

Between-task interference in probability distortion

If separate representations had been used for JRF and DMR, we would expect little interference of one task with the other. Instead, we found two-way interferences between the two tasks, which suggests shared representations of probability across tasks.

On one hand, the numerosity effect of probability distortion in JRF was modulated by DMR. As in Experiments 1 and 2, we fit the reported relative frequency in JRF to LLO and estimated $\hat{\gamma}$ and $\hat{p}_0$ for each participant and numerosity (Supplemental Fig. S6b). A repeated-measures ANOVA of numerosity was performed separately for $\hat{\gamma}$ and $\hat{p}_0$, and for Experiment 4A and 4B. In Experiment 4B, where JRF and DMR trials were blocked, numerosity had a significant effect on $\hat{\gamma}$ (F(4, 92) = 9.61, P < 0.0001, $\eta_p^2 = 0.29$), with $\hat{\gamma}$ at numerosity 200 significantly greater than that of numerosity 300, 400, 500, or 600 (Tukey's HSD post-hoc tests). In contrast, little numerosity effect on $\hat{\gamma}$ was found for Experiment 4A, where JRF and DMR trials were interleaved (F(4, 200) = 0.24, P = 0.91, $\eta_p^2 = 0.0048$). Numerosity also had a marginally significant effect on $\hat{p}_0$ in Experiment 4B (F(4, 92) = 2.39, P = 0.056, $\eta_p^2 = 0.094$), with $\hat{p}_0$ increasing with numerosity, but little effect in Experiment 4A (F(4, 200) = 0.91, P = 0.46, $\eta_p^2 = 0.018$). That is, the numerosity effect in Experiment 4 was parallel to those of Experiments 1 and 2 when JRF and DMR trials were blocked but not when they were interleaved.

On the other hand, the probability distortion in DMR seemed to be modulated by JRF. For a specific gamble $(x_1, p_1; x_2, p_2)$ and its certainty equivalent (CE) measured in the experiment, we defined the normalized CE[8] as $(CE - x_2)/(x_1 - x_2)$. In the framework of

[8] A similar normalized CE was defined in Tversky and Kahneman (1992) as CE/x_1, which only applies to gambles with $x_2 = 0$. Our definition can be considered as a generalization of theirs that allows $x_2 \neq 0$.

Cumulative Prospect Theory, we have $CE^{\alpha} = \pi(p_1)x_1^{\alpha} + (1-\pi(p_1))x_2^{\alpha}$, where α denotes the exponent of the power utility function. If $\alpha = 1$, the normalized CE would equal the subjective probability weight $\pi(p_1)$. As soon as $x_2 \le CE \le x_1$, the normalized CE is between 0 and 1. Thus the normalized CE could serve as a rough measure of $\pi(p_1)$. We computed the mean normalized CE across gambles for each participant and each p_1 (Supplemental Fig. S6c). As expected, the normalized CE was an inverted-*S*-shaped function of p_1. However, the normalized CE differed with task order: It was the highest in the DMR-first condition, medium in the two-task-interleaved condition, and lowest in the JRF-first condition. According to an 11 (p_1) by 3 (task order) mixed-design ANOVA on the normalized CE, there was a main effect of p_1 (F(10, 720) = 243.98, P < 0.0001, $\eta_p^2 = 0.77$), a marginally significant main effect of task order (F(2, 72) = 2.52, P = 0.087, $\eta_p^2 = 0.065$), and no interaction effect (F(20, 720) = 0.71, P = 0.82, $\eta_p^2 = 0.019$).

The interference between JRF and DMR in probability distortion provides evidence for a common representation of probability. We will see further and more direct evidence in the next section concerning invariance across tasks.

Replication and extension of the model comparison results of Experiments 1–3

In our further analysis, we collapsed data across Experiment 4A and 4B whenever possible. The JRF task of Experiment 4 is comparable to Experiments 1 and 2, while the DMR task is comparable to Experiment 3. For both tasks, we found BLO outperformed all alternative models in model comparisons (Fig. 4de and Supplemental Fig. S4). The assumptions of BLO in all three dimensions—log-odds representation, boundedness, and variance compensation—beat their alternative assumptions (Table 1). Figure 5 plots $\hat{\pi}(p) - p$ and

normalized CE as functions of p, contrasting the BLO and LLO predictions with data for individual participants. There was an evident improvement from LLO to BLO in capturing the observed probability distortion patterns.

Figure 5 about here

In Experiments 1–3, we had noticed that a few participants exhibited an *S*-shaped distortion, though most participants had the typical inverted-*S*-shaped distortion (Fig. 3 and Supplemental Fig. S3). The large number of participants (75 participants) in Experiment 4 allowed us to examine whether both types of participants can be modeled by BLO. In particular, we divided the participants into groups of $\hat{\gamma} \leq 1$ and those of $\hat{\gamma} > 1$, based on the $\hat{\gamma}$ estimated from the LLO model, separately for JRF and DMR. The BLO model outperformed all alternative models for both types of participants (Supplemental Fig. S5).

3. Invariances across Tasks and Individuals

Estimated parameters of BLO

What aspects of probability distortion may change with task or individual? What aspects are invariant? Does probability distortion reflect efficient coding, as we formulized in the Maximum Mutual Information hypothesis? We will examine the estimated parameters of BLO in all four experiments to address these questions.

The estimated parameters of BLO, summarized by their median values across participants, are summarized in Supplemental Table S2. We focused on the four key parameters of BLO that are not task-specific: Δ^-, Δ^+, τ, and Λ_0. As described earlier (Fig. 2), the lower bound Δ^- and upper bound Δ^+ define the range of input in log-odds that the brain chooses to represent at the moment. Any value outside the range cannot be precisely represented but cut off to its closest boundary. The scaling parameter τ maps this input range to the Thurstone scale. The anchor Λ_0 is the fixed point on the log-odds scale to expand or contract around in the compensation of encoding variance.

Invariance across individuals: the Thurstone constant

We computed Spearman's correlation (denoted r_s, which is robust to outliers) between the four BLO parameters across participants in the four experiments, separately for the JRF and DMR tasks. When Bonferroni corrections apply, Bonferroni corrected P values (denoted P_{BC}) are reported. Three pairs of parameters had significant correlations (Bonferroni corrected for 6 comparisons) that were consistent across tasks: Δ^- and Δ^+ (JRF: r_s = –0.87, P_{BC} < 0.0001; DMR: r_s = –0.33, P_{BC} = 0.010), Δ^- and τ (JRF: r_s = 0.72, P_{BC} < 0.0001; DMR: r_s = 0.75, P_{BC} < 0.0001), Δ^+ and τ (JRF: r_s = –0.78, P_{BC} < 0.0001; DMR: r_s = –0.46, P_{BC} < 0.0001).

As shown in the plots of Δ^+ against $-\Delta^-$ (Fig. 6a), the negative correlation between Δ^- and Δ^+ was a consequence of their near-symmetry around 0, that is, $\Delta^+ \approx -\Delta^-$. In other words, the bounds were symmetric around 0.5 on the probability scale. The symmetry may reflect the fact that p is always accompanied by $1-p$: for JRF, when the relative frequency of one color of dots is p, the other color is $1-p$; for DMR, when the probability for one outcome is p, the other outcome is $1-p$.

Figure 6 about here

The negative correlation between Δ^- and τ, and positive correlation between Δ^+ and τ suggest that τ may be inversely proportional to the half-range we defined in the Introduction,

$$\Delta = \left(\Delta^+ - \Delta^-\right)/2 \tag{21}$$

Indeed, we found that $1/\tau$ was highly positively correlated with Δ for both JRF (r_s = 0.73, P < 0.0001) and DMR (r_s = 0.71, P < 0.0001). That is, the product of τ and Δ was close to a constant across participants, whose median estimate was 3.89 for JRF and 2.03 for DMR (Fig. 6b). We call $\Psi \equiv \tau\Delta$ the Thurstone constant, the observed invariance of which suggests that the finite length of the internal Thurstone scale is similar for different individuals.

The Thurstone constant across time and tasks

One may doubt the invariance of $\tau\Delta$ might be a consequence of model redundancy that is determined by model structure. Instead, if Thurstone constant Ψ is a personal signature, we should expect it to vary slightly from individual to individual but be consistent for the same

individual across time and tasks. That is, there should be a positive correlation of Ψ across time and tasks.

In Experiment 4A, all participants completed two sessions on two different days, which allowed us to evaluate the consistency of Thurstone constant across time. As expected, we found significant positive correlation between Session 1's and Session 2's Ψ (Supplemental Fig. S7) for both JRF (r_s = 0.80, one-tailed P < 0.0001) and DMR (r_s = 0.57, one-tailed P < 0.0001).

In Experiment 4, we also found significant positive correlation between a participant's Ψ in JRF and the participant's Ψ in DMR (Fig. 6c): r_s = 0.21, one-tailed P = 0.039. This provides further evidence for BLO's hypothesis that a common cognitive constraint underlies probability representation in different tasks.

4. Maximizing Mutual Information

In all four experiments, participants had a bounded log-odds representation $\left[\Delta^{-},\Delta^{+}\right]$ that corresponds to a probability range far narrower than the range of objective probabilities: the bounded subjective ranges were [0.22, 0.79], [0.16, 0.86], [0.18, 0,90], [0.17, 0.81], and [0.16, 0.83], while the objective ranges were [0.01, 0.99], [0.02, 0.98], [0.01, 0.99], [0.01, 0.99], and [0.01, 0.99], respectively for Experiments 1, 2, 3, 4 (JRF), and 4 (DMR). As we will see below, this choice of $\left[\Delta^{-},\Delta^{+}\right]$ conforms with the Maximum Mutual Information hypothesis.

For our purposes, the efficiency of coding can be quantified by the mutual information between stimuli $s_1,\cdots,s_n$ and the corresponding responses $r_1,\cdots,r_n$:

$$I_m = \sum_{i=1}^{n} P\left(s_i,r_i\right)\log_2 \frac{P\left(s_i,r_i\right)}{P\left(s_i\right)P\left(r_i\right)}, \tag{22}$$

where $P\left(s_i\right)$ denotes the probability of occurrence of a specific stimulus s_i, $P\left(r_i\right)$ denotes the probability of occurrence of a specific response r_i, and $P\left(s_i,r_i\right)$ denotes the conjoint probability of the co-occurrence the two. Stimuli and responses refer to objective and subjective relative frequencies in JRF and refer to expected values and certainty equivalents in DMR. Mutual information is the Shannon information about the occurrence of the $s_1,...,s_n$ conveyed by the observed encoding. When stimuli and responses are independent of each other, i.e. $P\left(s_i,r_i\right)=P\left(s_i\right)P\left(r_i\right)$ then the mutual information $I_m=0$. Intuitively, when probability and encoding are statistically independent, knowing one tells us nothing about the other.

We were interested in how the expected mutual information may vary with Δ^{+} and Δ^{-}. For a specific experiment, we used the BLO model to generate simulated responses for the stimuli of the experiment and then computed the expected mutual information for different Δ^{+}

and Δ^- using a Monte Carlo method. To obtain a stable estimate of the expected mutual information, we repeated the stimulus set of each experiment to produce 99,000 trials for JRF and 330,000 trials for DMR. The parameters used for response generation were the median of the estimated parameters and the median Ψ, except that the value of τ was determined by the invariance $\tau = \Psi/\Delta$. In the computation of mutual information, for JRF, the objective and subjective probabilities were rounded to the 2nd decimal, and for DMR, the expected values and certainty equivalents were rounded to the closest integer.

In all four experiments, the expected mutual information varied non-monotonically with the values of Δ^+ and Δ^- (Fig. 7) and there is evidently a unique maximum over the range considered. For all four experiments, the observed median values of Δ^+ and Δ^- (marked by solid red circles) were close to the values maximizing the expected mutual information: the mutual information associated with the observed Δ^+ and Δ^- were only 4.12%, 5.65%, 1.34%, 2.68%, and 0.16% lower than maximum, respectively for Experiments 1, 2, 3, 4 (JRF), and 4 (DMR).

Figure 7 about here
---

5. Discussion

In decision under risk, participants act as if their choices were based on probability values systematically different from those that are objectively correct. Similar systematic distortion is found in tasks involving relative frequency judgments and confidence ratings (Zhang & Maloney, 2012). The distortions are *dynamic*: the same participant can have different distortions in different tasks (Wu, Delgado & Maloney, 2009). We proposed a model of probability distortion as a form of bounded rationality—the Bounded Log-Odds (BLO) model—and test it in four experiments. We consider distortion in judgments of relative frequency (JRF) and decision under risk (DMR).

BLO is based on a small number of assumptions commonly found in cognitive and perceptual models. We independently tested each of these assumptions (log-odds, boundedness, variance compensation, etc.) using factorial model comparison to verify that they are all essential to fitting human data. They are. If we replace any assumption by the alternatives we consider, the resulting model is strictly inferior to BLO.

We further show that human performance comes close to maximizing **the mutual information** between objective probabilities in a task and their imperfect internal representations. In the four experiments we report, participants chose probability distortions consistent with BLO and also with maximizing mutual information.

A recent article (Polanía et al., 2019) uses the same criterion (maximum mutual information) to model human encoding of value. The two articles taken together are consistent with a claim, supported by considerable experimental data, that many observed failures in DMR can be viewed as attempts to compensate for immutable limits in cognitive processing in order to preserve Shannon information, a form of bounded rationality (Simon, 1982).

There are many theoretical models intended to account for inverted-*S*- or *S*-shaped probability distortion: the power model of proportion judgment (Hollands & Dyre, 2000; Spence, 1990), the support theory model of probability judgment (Fox & Rottenstreich, 2003; Fox & Tversky, 1998), the calibration model (Smith & Ferrell, 1983), the stochastic model of confidence rating (Erev et al., 1994; Wallsten et al., 1997), and the adaptive probability theory model of decision under risk (Martins, 2006).

Almost all these models were proposed for one specific type of task and are not intended as general explanations for probability distortion. Neither do they explain the variation of probability distortion across task settings—there was even a belief, at least in decision under risk, that the parameters of distortion should be the same for all gambles (Birnbaum, 2005).

In contrast, BLO characterizes a common mechanism underlying all probability distortion, where we identified two constraints—boundedness and compensation for representational uncertainty (variance)—that are pervasive in models of cognitive and perceptual tasks (Nassar, Helmers, & Frank, 2018; Orhan & Jacobs, 2013; van den Berg et al., 2014; Wei & Stocker, 2015). BLO can be used to estimate an individual's probability distortion in one task and predict the same individual's performance in another task. BLO can account for how the estimated slope may vary with tasks or different settings of the same task. We next describe some of the implications of BLO.

Discontinuities at p = 0 and p = 1

BLO and any model based on the boundedness assumption predict that $\pi(0) > 0$ and $\pi(1) < 1$, that is, probability distortion with discontinuities at $p = 0$ and $p = 1$ (e.g. see the $\pi(p)$ curves in Supplemental Fig. S3). Such discontinuities are also found in the neo-additive family of weighting functions (Wakker, 2010), but not in widely accepted forms of probability

distortion such as LLO (Gonzalez & Wu, 1999; Zhang & Maloney, 2012) and Prelec's function (Prelec, 1998). Kahneman and Tversky had allowed a similar discontinuity in their original settings for probability weighing functions (Figure 4 of Kahneman & Tversky, 1979). Their later shift (Tversky & Kahneman, 1992) to continuous probability functions seems motivated by a desire to have simple mathematical representations of probability distortion.

The ranges between the estimated bounds that we found are fairly limited, approximately [0.15, 0.85]. Given that the occurrence of probabilities as extreme as 0.05 and 0.95—or even 0.01 and 0.99—is not uncommon in laboratory tasks or real life, bounding is likely to exert detectable influences on probability representation and performance under many circumstances. Indeed, there are clues indicating boundedness in previous studies. For example, Yang and Shadlen (2007) studied monkeys' probabilistic inference and found that the strength of a specific evidence perceived by the monkey was, in general, proportional to the objective log-odds of the evidence. But for "sure evidence" that corresponded to minus or plus infinity in log-odds, the subjective log-odds were bounded, equivalent to [0.15, 0.81] and [0.30, 0.64] in probability for the two tested monkeys.

Compensation for encoding uncertainty

Anchoring—as a way to explain the inverted-*S*-shaped curve and its individual differences—has been assumed in a few theories or models of probability distortion (Fox & Rottenstreich, 2003; Martins, 2006; See, Fox, & Rottenstreich, 2006; see Zhang & Maloney, 2012 for a summary). It can be a way to improve the accuracy of probability judgment, following the perspective of Bayesian inference (Petzschner et al., 2015). What distinguishes BLO from previous models is the assumption that anchoring implements compensation for encoding uncertainty. Intuitively, percepts of lower uncertainty should be less discounted and those of higher uncertainty more discounted. If the uncertainty of a percept varies with the value of probability it encodes, so will the reliability weight endowed to the percept.

For the JRF task, the uncertainty may arise from a sampling process, analogous to the sampling in perceptual tasks such as motion perception (Dakin, Mareschal, & Bex, 2005; Raidvee, Averin, Kreegipuu, & Allik, 2011) and pattern detection (B. F. Green, Wolf, & White, 1959). Compensation for sampling variation enables the pattern of probability distortion to vary with task settings for the same individual and accounts for the numerosity effect observed in JRF.

For the DMR task, where probability is explicitly defined and the sampling process does not seem to apply, we still found that the slope of probability distortion relies on a $p(1-p)$ term, varying with p. It is as if people are compensating for the variation of a virtual sampling process, or the variation caused by Gaussian noise on the Thurstonian log-odds scale (Appendix A, see also Lebreton, Abitbol, Daunizeau, & Pessiglione, 2015). Lebreton et al. (2015) show that a generalized form of $p(1-p)$ is correlated with the confidence of value or probability perception and is automatically encoded in the ventromedial prefrontal cortex (vmPFC) of the human brain. Under certain circumstances, such non-linear anchoring may result in counterintuitive non-monotonic probability distortion, as we elaborate below.

Predicted violations of stochastic dominance

We would like to highlight a counterintuitive prediction of BLO that agrees with empirical observations, the *violation of stochastic dominance* observed for one participant in the decision-under-risk experiment. A probability weighting function, $\pi(p)$, predicted by LLO would be monotonically increasing with the objective probability, but that of BLO need not be, as demonstrated by the $\pi(p)$ of Participant 4 in Experiment 3 (Supplemental Fig. S3c, thickened light blue curve). The possible non-monotonicity comes from the non-linear transformation that compensates for encoding variance (Eq. 6), where the weight for the

anchor depends on the value of the objective probability, and may be exaggerated by boundedness. For the participant, the observed non-monotonicity in $\pi(p)$ agrees with her consistent violation of stochastic dominance in specific risky choices.

The violation of stochastic dominance in Participant 4's choice behaviors was subtle, going unnoticed in the original paper (Gonzalez & Wu, 1999). We noticed it only because the BLO model fitted to this participant predicted such violations. There may well be more such participants overlooked in previous studies.

The BLO model also permits a more dramatic form of violation of stochastic dominance, one that is also reported in the literature: given a choice between a sure reward and a lottery with probability $p < 1$ of receiving the same reward and otherwise nothing, some participants choose the lottery more than 50% of the time (Blankenstein, Crone, van den Bos, & van Duijvenvoorde, 2016; Tymula et al., 2012). Their choices implied that $\pi(1) < \pi(p)$, that is, $\pi(p)$ is not an increasing (more precisely, non-decreasing) function of p. In past work, such participants were typically excluded from further analyses. As we saw above, these participants' violation of stochastic dominance is predicted by BLO for certain parameter settings that result in a non-monotonic $\pi(p)$.

Birnbaum and colleagues' elegant transfer-of-attention-exchange model (Birnbaum, 2005; Birnbaum & Navarrete, 1998) predicted the violations of stochastic dominance on the group level in a series of gambles. However, for gambles with only two branches, the model is reduced to a variant of CPT whose probability weighting function $\pi(p)$ varies monotonically with p (Appendix D). As the result, it cannot explain the violation of stochastic dominance in our case. In other words, the BLO model and the transfer-of-attention-exchange model account for violations of stochastic dominance under different circumstances.

Predicting the slope of the probability distortion function

A second prediction of BLO concerns *the control of the slope of probability distortion*. We found invariances between Δ^-, Δ^+, and τ, which predicts that the slope of probability distortion would be partly controlled by the probability range of stimuli: The narrower the probability range of stimuli, the greater the slope of distortion. As a special case, when the probability range of stimuli is narrower than the output range, participants should tend to underestimate small probabilities and overestimate large probabilities (though not necessarily so, due to the anchoring effect we discussed earlier). That is, participants are likely to show an atypical *S*-shaped (instead of inverted-*S*-shaped) probability distortion, or in other words, a greater-than-one slope of distortion ($\gamma > 1$). Indeed, for all the relative frequency judgment studies we cited earlier where an *S*-shaped distortion was found (Brooke & MacRae, 1977; Pitz, 1966; Shuford, 1961), the range of stimuli—$\left[\ 0.1, 0.9\ \right]$was relatively narrow. In contrast, those who used a similar task but reported an inverted-*S*-shaped distortion (Erlick, 1964; Varey et al., 1990) usually used a range as wide as $\left[\ 0.02, 0.98\ \right]$.

For decision under risk, the following meta-analysis based on previous empirical studies supports our prediction of how the slope of distortion may vary with the probability range of stimuli. Fox and Poldrack (2014, Table A.3) summarized the results of a number of decision-making studies that were modeled in the framework of Prospect Theory. In Fox and Poldrack's list, we identified the studies where the gamble set was explicitly defined and each gamble consisted of two outcomes that could be denoted $\left(x_1, p; x_2, 1-p\right)$ (see Supplemental Table S3 for the 12 studies included). Though different functional forms—LLO, Prelec's (1998) one-parameter and two-parameter functions, and Tversky and Kahneman's (1992) weighting function—had been assumed in different studies, all had a parameter for the slope of

probability distortion that is roughly equivalent to the γ in LLO[9]. For each study, we computed the standard deviation of $\lambda(p)$'s distribution as a measure of the probability range of stimuli. As predicted by BLO, we found this measure[10] was significantly negatively correlated with the slope of probability distortion (Fig. 8), Spearman's r_s = –0.65, one-tailed P = 0.010.

Figure 8 about here

Common cognitive constraints for probability representation

One may wonder whether the pattern of probability distortion observed in the JRF task, such as the numerosity effect, was simply caused by factors specific to the visual modality—given that the area of display was constant, two displays that differed in the numerosity of dots also differed in the visual density of dots. This modality-specificity is unlikely to be true. On one hand, the estimation of numerosity is little influenced by density (Dakin, Tibber, Greenwood, Kingdom, & Morgan, 2011), especially when area is constant and density covaries with numerosity, as in our case. On the other hand, the properties (e.g. boundedness, compensation for sampling variation) found in the JRF experiments were also found in the

[9] Among these studies a few provided slope parameters estimated from multiple weighting functions: 0.67 and 0.61 for Glöckner and Pachur (2012), 0.96, 1.00, 0.94, and 0.96 for Stott (2006). We can see the slope parameters estimated from different probability weighting functions were almost the same. In our analysis, when multiple slope parameters had been reported for a study, we used their mean. Since not all studies involved the loss domain, we only considered the slope of distortion for the gain domain, that is, the γ^+ in Fox and Poldrack (2014).

[10] When computing the standard deviation of $\lambda(p)$ for a specific study, we took into account the probabilities of both outcomes in each gamble, that is, both p and $1-p$ in $(x_1, p; x_2, 1-p)$. If only the probability for the better outcome, p, were used for the measure, the resulting correlation with the slope of probability distortion would be lower: Spearman's r_s = –0.56, one-tailed P = 0.030.

DMR experiments, suggesting cognitive underpinnings that go beyond a single modality or task. Even if the probability distortion observed in the visual judgment task had visual causes, it would not change the fact that BLO and its assumptions describe well the empirical data in the two different tasks.

According to BLO, a bounded range $\left[\Delta^{-},\Delta^{+}\right]$ of the log-odds scale is mapped to a Thurstone scale of fixed length by a scaling factor τ. Indeed, for both JRF and DMR, we found $\tau\Delta$, which corresponds to the half length of the Thurstone scale, was almost invariant across participants. We defined $\Psi \equiv \tau\Delta$ and termed it the Thurstone constant.

Though individual differences in Ψ are small, an individual's Ψ at one time point is positively correlated with the individual's Ψ at another time point. The individual's Ψ is also positively correlated across tasks. Thus the Thurstone constant proves to be a personal signature that imposes a common constraint on the representation of probability distortion in different tasks.

Open questions and future directions

The judgment of relative frequency and decision under risk are the only two tasks where BLO and its assumptions have been tested, but these two tasks together represent a vast body of previous research. The model may be applied to a wider range of tasks involving frequency and probability. It will likely shed light on the common and distinctive mechanisms of probability distortion in different tasks.

What determine the slope and crossover point of probability distortion in a specific task? Why may the parameters of probability distortion change from task to task and from individual to individual? In the present study we have provided tentative answers. A full answer to these questions awaits research in the future.

A puzzle we did not address earlier concerns the crossover point of probability distortion (i.e. the point on the distortion curve where overestimation changes into underestimation or the reverse). It has been frequently observed that the crossover point is near 0.5 for the JRF task (Zhang & Maloney, 2012) but approximately 0.37 for the DMR task (Prelec, 1998). That is, the probability distortion is symmetric around 0.5 in the former but asymmetric in the latter. There are plausible reasons to have symmetry, but why asymmetry? Here we conjecture that the asymmetry is also driven by the maximization of mutual information, which, for the DMR task, means to have the CEs of different gambles as discriminable as possible. Following conventions (Fox & Poldrack, 2014; Gonzalez & Wu, 1999) and for parsimony, we had assumed a uniform Gaussian noise on the CE scale. However, larger CEs may tend to be associated with higher variances, according to Weber's law (Gibbon, 1977). To compensate for this, more of the representational scale should be devoted to larger probabilities and thus to the larger CEs associated with them. Indeed, adopting a smaller crossover point (i.e. less than 0.5) would map larger probabilities to a longer range of subjective probabilities and effectively implement such a strategy of probability representation.

Important questions for future research also include: How may probability distortion change from trial to trial? We conjecture that the human representation of probability can adapt to the environment, in the spirit of efficient coding (Attneave, 1954; Barlow, 1961; Simoncelli & Olshausen, 2001). The current version of BLO is a stationary model, whose prediction will not change with time or experience. In contrast, non-stationarity has been identified in probability distortion for both the judgment of relative frequency (Zhang & Maloney, 2012) and decision under risk (Zeisberger, Vrecko, & Langer, 2012).

Can the BLO model apply to decisions among complex gambles with more than two non-zero outcomes? In theory, BLO just specifies the probability weighting function in Prospect Theory and can apply to any circumstances where Prospect Theory applies. But it is

still an empirical question whether probability distortion for gambles with more than two non-zero outcomes can be predicted by BLO.

A final note: Kahneman and Tversky's original Prospect Theory contained the assumption that decision makers would first interpret ("edit") available information (Kahneman & Tversky, 1979). In this initial *editing* stage they might, for example, convert the probability 0.31317 to the more tractable 1/3. Only then would they assign prospect values to lotteries in the second, *evaluation* stage. In presenting the BLO Model we focus on evaluation. Still, nothing about the theory would preclude adding an editing phase or discretizing the representation of probability if justified by empirical results.

References

Abdellaoui, M., Diecidue, E., & Öncüler, A. (2011). Risk Preferences at Different Time Periods: An Experimental Investigation. *Management Science, 57*(5), 975-987. doi:10.1287/mnsc.1110.1324

Abdellaoui, M., L'Haridon, O., & Paraschiv, C. (2011). Experienced vs. Described Uncertainty: Do We Need Two Prospect Theory Specifications? *Management Science, 57*(10), 1879-1895. doi:10.1287/mnsc.1110.1368

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19*(6), 716-723. doi:10.1109/TAC.1974.1100705

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine (The behavior of a rational agent in the face of risk: Critique of the postulates and axioms of the American school). *Econometrica: Journal of the Econometric Society, 21*(4), 503-546.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, Illinois: Waveland Press.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*(3), 183-193.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217-234). Cambridge, MA: MIT Press.

Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society. Series B., 11*(2), 115-149.

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making, 16*(3), 215-233. doi:10.1002/bdm.443

Bell, D. E. (1985). Disappointment in decision making under uncertainty. *Operations Research, 33*(1), 1-27.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association, 39*(227), 357-365.

Bernoulli, D. (1738/1954). Exposition of a new theory on the measurement of risk. *Econometrica, 22*(1), 23-36.

Birnbaum, M. H. (1992). Violations of monotonicity and contextual effects in choice-based certainty equivalents. *Psychological Science, 3*(5), 310-315.

Birnbaum, M. H. (2005). Three New Tests of Independence That Differentiate Models of Risky Decision Making. *Management Science, 51*(9), 1346-1358. doi:10.1287/mnsc.1050.0404

Birnbaum, M. H., & Navarrete, J. B. (1998). Testing descriptive utility theories: Violations of stochastic dominance and cumulative independence. *Journal of Risk and Uncertainty, 17*, 49–78.

Birnbaum, M. H., & Stegner, S. E. (1981). Measuring the importance of cues in judgment for individuals: Subjective theories of IQ as a function of heredity and environment. *Journal of Experimental Social Psychology, 17*(2), 159–182.

Blackwell, D., & Girshick, M. A. (1954). *Theory of games and statistical decisions*. New York: Wiley.

Blankenstein, N. E., Crone, E. A., van den Bos, W., & van Duijvenvoorde, A. C. (2016). Dealing with uncertainty: testing risk-and ambiguity-attitude across adolescence. *Developmental Neuropsychology, 41*(1-2), 77-92. doi:10.1080/87565641.2016.1158265

Brooke, J. B., & MacRae, A. W. (1977). Error patterns in the judgment and production of numerical proportions. *Perception and Psychophysics, 21*(4), 336-340.

Bruhin, A., Fehr-Duda, H., & Epper, T. (2010). Risk and Rationality: Uncovering Heterogeneity in Probability Distortion. *Econometrica, 78*(4), 1375-1412. doi:10.3982/ecta7139

Burr, D., & Ross, J. (2008). A Visual Sense of Number. *Current Biology, 18*(6), 425-428. doi:10.1016/j.cub.2008.02.052

Cavagnaro, D. R., Pitt, M. A., Gonzalez, R., & Myung, J. I. (2013). Discriminating Among Probability Weighting Functions Using Adaptive Design Optimization. *Journal of Risk and Uncertainty, 47*(3), 255-289. doi:10.1007/s11166-013-9179-3

Chateauneuf, A., Eichberger, J., & Grant, S. (2007). Choice under uncertainty with the best and worst in mind: Neo-additive capacities. *Journal of Economic Theory, 137*(1), 538-567.

Cicchini, G. M., Anobile, G., & Burr, D. C. (2014). Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proceedings of the National Academy of Sciences*, 201402785. doi:10.1073/pnas.1402785111

Cochran, W. G. (1977). *Sampling techniques.* New York, NY: John Wiley & Sons.

Cohen, M. (1992). Security level, potential level, expected utility: A three-criteria decision model under risk. *Theory and Decision, 33*(2), 101-134.

Dakin, S. C., Mareschal, I., & Bex, P. J. (2005). Local and global limitations on direction integration assessed using equivalent noise analysis. *Vision Research, 45*(24), 3027-3049. doi:10.1016/j.visres.2005.07.037

Dakin, S. C., Tibber, M. S., Greenwood, J. A., Kingdom, F. A. A., & Morgan, M. J. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences of the United States of America, 108*(49), 19552-19557.

Daunizeau, J. (2017). *Semi-analytical approximations to statistical moments of sigmoid and softmax mappings of normal variables*. https://sites.google.com/site/jeandaunizeauswebsite/.

DeBruine, L. M., Smith, F. G., Jones, B. C., Roberts, S. C., Petrie, M., & Spector, T. D. (2009). Kin recognition signals in adult faces. *Vision Research, 49*, 38-43.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., . . . Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making, 23*(1), 15-47.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*(3), 519-527.

Erlick, D. E. (1964). Absolute judgments of discrete quantities randomly distributed over time. *Journal of Experimental Psychology, 67*(5), 475-482.

Fennell, J., & Baddeley, R. (2012). Uncertainty plus prior equals rational bias: An intuitive Bayesian probability weighting function. *Psychological Review, 119*(4), 878-887.

Fox, C. R., & Hadar, L. (2006). Decisions from experience = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making, 1*(2), 159-161.

Fox, C. R., & Poldrack, R. A. (2014). Prospect theory and the brain. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics* (2nd ed., pp. 533–567). London, UK: Academic Press.

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science, 14*(3), 195-200.

Fox, C. R., & Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science, 44*(7), 879-895.

Gächter, S., Johnson, E. J., & Herrmann, A. (2007). Individual-level loss aversion in riskless and risky choices. *SSRN eLibrary*.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review, 84*, 279–325. doi:10.1037/0033-295X.84.3.279

Gilboa, I. (1988). A combination of expected utility and maxmin decision criteria. *Journal of Mathematical Psychology, 32*(4), 405-420.

Gilchrist, A., Kossyfidis, C., Bonato, F., Agostini, T., Cataliotti, J., Li, X., . . . Economou, E. (1999). An anchoring theory of lightness perception. *Psychological Review, 106*(4), 795-834. doi:10.1037/0033-295x.106.4.795

Glöckner, A., & Betsch, T. (2008). Do people make decisions under risk based on ignorance? An empirical test of the priority heuristic against cumulative prospect theory. *Organizational Behavior and Human Decision Processes, 107*(1), 75-95. doi:10.1016/j.obhdp.2008.02.003

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition, 123*(1), 21-32. doi:10.1016/j.cognition.2011.12.002

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences, 5*(1), 10-16.

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron, 36*, 299-308.

Goldstein, W. M., & Einhorn, H. J. (1987). Expression theory and the preference reversal phenomena. *Psychological Review, 94*(2), 236-254.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology, 38*(1), 129-166. doi:10.1006/cogp.1998.0710

Green, B. F., Wolf, A. K., & White, B. W. (1959). The Detection of Statistically Defined Patterns in a Matrix of Dots. *American Journal of Psychology, 72*(4), 503-520.

Green, D. M., & Swets, J. A. (1966/1974). *Signal detection theory and psychophysics*. New York: Wiley.

Harrison, G. W., & Rutström, E. E. (2009). Expected utility theory and prospect theory: one wedding and a decent funeral. *Experimental Economics, 12*(2), 133–158. doi:10.1007/s10683-008-9203-7

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534-539. doi:10.1111/j.0956-7976.2004.00715.x

Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review, 107*, 500-524.

Holt, C. A., & Laury, S. K. (2002). Risk Aversion and Incentive Effects. *American Economic Review, 92*(5), 1644-1655. doi:10.1257/000282802762024700

Hsu, M., Krajbich, I., Zhao, C., & Camerer, C. F. (2009). Neural response to reward anticipation under risk is nonlinear in probabilities. *The Journal of Neuroscience, 29*(7), 2231-2237.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297-307. doi:10.1093/biomet/76.2.297

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-291.

Knight, F. H. (1921). *Risk, uncertainty and profit*. Boston, MA: Houghton Mifflin.

Kobayashi, S., Carvalho, O. P. D., & Schultz, W. (2010). Adaptation of reward sensitivity in orbitofrontal neurons. *Journal of Neuroscience, 30*, 534–544.

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience, 18*, 1159–1167. doi:10.1038/nn.4064

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 551-578.

Loomes, G., Moffatt, P. G., & Sugden, R. (2002). A microeconometric test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty, 24*(2), 103-130.

Louie, K., Khaw, M. W., & Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences, 110*(15), 6139-6144. doi:10.1073/pnas.1217854110

Luce, R. D. (2000). *Utility of gains and losses: Measurement-theoretical and experimental approaches*. London: Lawrence Erlbaum.

Maloney, L. T., & Dal Martello, M. F. (2006). Kin recognition and the perceived facial similarity of children. *Journal of Vision, 6*(10), 1047-1056.

Maloney, L. T., & Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Research, 50*(23), 2362-2374. doi:doi: 10.1016/j.visres.2010.09.031

Martins, A. C. R. (2006). Probability biases as Bayesian inference. *Judgment and Decision Making, 1*(2), 108-117.

Mellers, B., Weiss, R., & Birnbaum, M. (1992). Violations of dominance in pricing judgments. *Journal of Risk and Uncertainty, 5*(1), 73-90.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a Rational Strategy for Lossy Data Compression in Visual Working Memory. *Psychological Review, 125*(4), 486–511. doi:10.1037/rev0000101

Oberauer, K., & Lin, H.-Y. (2017). An Interference Model of Visual Working Memory. *Psychological Review, 124*(1), 21–59. doi:10.1037/rev0000044

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review, 120*(2), 297–328.

Oruç, I., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research, 43*(23), 2451-2468.

Payzan-LeNestour, E., Balleine, Bernard W., Berrada, T., & Pearson, J. (2016). Variance After-Effects Distort Risk Perception in Humans. *Current Biology, 26*(11), 1500-1504. doi:10.1016/j.cub.2016.04.023

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences, 19*(5), 285-293.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72*(3), 346-354.

Pitz, G. F. (1966). The sequential judgment of proportion. *Psychonomic Science, 4*, 397-398.

Polanía, R., Woodford, M., & Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience, 22*(1), 134-142. doi:10.1038/s41593-018-0292-0

Prelec, D. (1998). The probability weighting function. *Econometrica, 66*, 497-527.

Raidvee, A., Averin, K., Kreegipuu, K., & Allik, J. (2011). Pooling elementary motion signals into perception of global motion direction. *Vision Research, 51*(17), 1949-1957. doi:10.1016/j.visres.2011.07.004

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1446-1465.

Rustichini, A., Conen, K. E., Cai, X., & Padoa-Schioppa, C. (2017). Optimal coding and neuronal adaptation in economic decisions. *Nature communications, 8*(1), 1208. doi:10.1038/s41467-017-01373-y

Savage, L. J. (1954). *The foundations of statistics.* New York: Wiley.

See, K. E., Fox, C. R., & Rottenstreich, Y. S. (2006). Between ignorance and truth: Partition dependence and learning in judgment under uncertainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(6), 1385-1402.

Shannon, C. E. (1948a). A mathematical theory of communication I. *Bell System Technical Journal, 27*, 379-423.

Shannon, C. E. (1948b). A mathematical theory of communication II. *Bell System Technical Journal, 27*, 623-656.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

Shuford, E. H. (1961). Percentage estimation of proportion as a function of element type, exposure time, and task. *Journal of Experimental Psychology, 61*(5), 430-436.

Simon, H. A. (1982). *Models of bounded rationality: Economic Analysis and Public Policy* (Vol. 1). Cambridge, MA: MIT press.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*(1), 1193-1216.

Smith, M., & Ferrell, W. R. (1983). The effect of base rate on calibration of subjective probability for true-false questions: Model and experiment. In P. Humphreys, O. Svenson, & A. Vari (Eds.), *Analysing and Aiding Decision Processes* (pp. 469-488). Amsterdam: North Holland.

Spence, I. (1990). Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance, 16*(4), 683-692.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological monographs: General and applied, 74*(11), 1-29.

Stauffer, W. R., Lak, A., Bossaerts, P., & Schultz, W. (2015). Economic Choices Reveal Probability Distortion in Macaque Monkeys. *The Journal of Neuroscience, 35*(7), 3146-3154. doi:10.1523/jneurosci.3653-14.2015

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage, 46*(4), 1004-1017. doi:10.1016/j.neuroimage.2009.03.025

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty, 32*(2), 101-130. doi:10.1007/s11166-006-8289-6

Tanaka, T., Camerer, C. F., & Nguyen, Q. (2010). Risk and Time Preferences: Linking Experimental and Household Survey Data from Vietnam. *American Economic Review, 100*(1), 557-571. doi:10.1257/aer.100.1.557

Teitelbaum, J. C. (2007). A unilateral accident model under ambiguity. *The Journal of Legal Studies, 36*(2), 431-477.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286.

Tobler, P. N., Christopoulos, G. I., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2008). Neuronal distortions of reward probability without choice. *The Journal of Neuroscience, 28*(45), 11703-11711.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive Coding of Reward Value by Dopamine Neurons. *Science, 307*(5715), 1642-1645. doi:10.1126/science.1105370

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2008). Decision making, movement planning and statistical decision theory. *Trends in Cognitive Sciences, 12*(8), 291-297.

Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review, 102*(2), 269-283.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297-323. doi:10.1007/BF00122574

Tymula, A., Belmaker, L. A. R., Ruderman, L., Glimcher, P. W., & Levy, I. (2013). Like cognitive function, decision making across the life span shows profound age-related changes. *Proceedings of the National Academy of Sciences, 110*(42), 17143-17148.

Tymula, A., Rosenberg Belmaker, L. A., Roy, A. K., Ruderman, L., Manson, K., Glimcher, P. W., & Levy, I. (2012). Adolescents' risk-taking behavior is driven by tolerance to ambiguity. *Proceedings of the National Academy of Sciences, 109*(42), 17135-17140. doi:10.1073/pnas.1207144109

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science, 20*(4), 473-479. doi:10.1111/j.1467-9280.2009.02319.x

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121*(1), 124–149. doi:10.1037/a0035234

Varey, C. A., Mellers, B. A., & Birnbaum, M. H. (1990). Judgments of proportions. *Journal of Experimental Psychology: Human Perception and Performance, 16*(3), 613-625.

von Neumann, J., & Morgenstern, O. (1944/1953). *Theory of games and economic behavior* (3rd ed.). Princeton, NJ: Princeton University Press.

Vrecko, D., & Langer, T. (2013). What Are Investors Willing to Pay to Customize Their Investment Product? *Management Science, 59*(8), 1855-1870. doi:10.1287/mnsc.1120.1695

Wackerly, D. D., Mendenhall, W. I., & Scheaffer, R. L. (2008). *Mathematical statistics with applications* (7th ed.). Belmont, CA: Thomson.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge, UK: Cambridge University Press.

Wald, A. (1949). Statistical Decision Functions. *The Annals of Mathematical Statistics, 20*(2), 165-205.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10*(3), 243-268.

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain ‘anti-Bayesian’ percepts. *Nature Neuroscience, 18*, 1509–1517. doi:10.1038/nn.4105

Wolpert, D. M., & Landy, M. S. (2012). Motor control is decision-making. *Current Opinion in Neurobiology, 22*(6), 996-1003.

Wu, S.-W., Delgado, M. R., & Maloney, L. T. (2009). Economic decision-making under risk compared with an equivalent motor task. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 6088-6093.

Wu, S.-W., Delgado, M. R., & Maloney, L. T. (2015). Motor decision-making. In A. Toga (Ed.), *Brain Mapping: An Encyclopedic Reference* (Vol. 3, pp. 417–427): Elsevier Science & Technology.

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature, 447*, 1075-1080. doi:10.1038/nature05852

Zeisberger, S., Vrecko, D., & Langer, T. (2012). Measuring the time stability of Prospect Theory preferences. *Theory and Decision, 72*(3), 359-386. doi:10.1007/s11238-010-9234-3

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action and cognition. *Frontiers in Neuroscience, 6*, 1. doi:10.3389/fnins.2012.00001

Appendix A: Derivation of the encoding variance

Suppose X is an encoding of the probability p on the log-odds scale contaminated by a Gaussian noise: $X \sim Normal\left(\lambda(p), \Sigma\right)$, where $\lambda(p) = \log\frac{p}{1-p}$ and Σ is a constant for variance. When transformed back to the probability scale, the representation is $\lambda^{-1}[X]$. Given that $\lambda^{-1}[X] = \frac{1}{1+e^{-X}}$ is a logistic function of the Gaussian random variable X, its variance has the following approximation (Daunizeau, 2017):

$$Var\left(\lambda^{-1}[X]\right) \approx \lambda^{-1}\left[\frac{\lambda(p)}{\sqrt{1+\frac{3\Sigma}{\pi^2}}}\right]\left(1-\lambda^{-1}\left[\frac{\lambda(p)}{\sqrt{1+\frac{3\Sigma}{\pi^2}}}\right]\right)\left(1-\frac{1}{\sqrt{1+\frac{3\Sigma}{\pi^2}}}\right). \tag{A1}$$

Let $\zeta = \frac{3\Sigma}{\pi^2}$, the right side of Eq. A1 can be written as a function of ζ:

$$g(\zeta) = \lambda^{-1}\left[\frac{\lambda(p)}{\sqrt{1+\zeta}}\right]\left(1-\lambda^{-1}\left[\frac{\lambda(p)}{\sqrt{1+\zeta}}\right]\right)\left(1-\frac{1}{\sqrt{1+\zeta}}\right). \tag{A2}$$

It is easy to prove that the values of $g(\zeta)$ and its first derivative at $\zeta = 0$ are $g(\zeta)\big|_{\zeta=0} = 0$ and $g'(\zeta)\big|_{\zeta=0} = \frac{1}{2}p(1-p)$. When $\zeta = \frac{3\Sigma}{\pi^2} \ll 1$, the Taylor expansion of $g(\zeta)$ on ζ is:

$$\begin{aligned} g(\zeta) &= g(\zeta)\big|_{\zeta=0} + g'(\zeta)\big|_{\zeta=0}\zeta + o\left(\zeta^2\right) \\ &= \frac{1}{2}p(1-p)\zeta + o\left(\zeta^2\right) \\ &\approx \frac{1}{2}p(1-p)\zeta = \frac{3\Sigma}{2\pi^2}p(1-p) \end{aligned} . \tag{A3}$$

That is, the variance of $\lambda^{-1}[X]$ has the approximate form,

$$Var\left(\lambda^{-1}\left[X\right]\right) \approx \frac{3\Sigma}{2\pi^2} p\left(1-p\right) . \tag{A4}$$

See Lebreton et al. (2015) for a similar proof in a different guise.

Appendix B: Derivation of the variance of sampling without replacement

Suppose that $B_1, \cdots, B_{n_s}$ are sampled[11] *without replacement* from a population $\{b_1, \cdots, b_N\}$ of $N \geq n_s$ members taking on only the values 0 or 1. The mean of the population is $p = N^{-1}\sum_{i=1}^{N} b_i$, the proportion of the population that takes on the value 1. The mean of the sample is $P_s = n_s^{-1}\sum_{i=1}^{n_s} B_i$ (a random variable) and the variance of P_s is (Cochran, 1977, Theorem 2.2):

$$
\begin{aligned}
Var(P_s) &= E[P_s - p]^2 \\
&= \frac{\sum_{i=1}^{N}(b_i - p)^2}{N-1}\frac{1}{n_s}\frac{N-n_s}{N} \\
&= \frac{\sum_{i=1}^{N} b_i^2 - 2p\sum_{i=1}^{N} b_i + \sum_{i=1}^{N} p^2}{n_s N}\frac{N-n_s}{N-1}. \\
&= \frac{Np - 2pNp + Np^2}{n_s N}\frac{N-n_s}{N-1} \\
&= \frac{p(1-p)}{n_s}\frac{N-n_s}{N-1}
\end{aligned}
\tag{A5}
$$

The first term $p(1-p)/n_s$ on the right hand side is the variance of samples drawn with replacement. The second term $(N-n_s)/(N-1)$ is the "correction for sampling without replacement". Eq. A5 can be derived alternatively by recognizing that NP_s, as a random variable, follows a hypergeometric distribution (Wackerly et al., 2008).

When $n_s = 1$, sampling with and without replacement are indistinguishable and the sample variance is the familiar

$$Var(B_1) = p(1-p) \tag{A6}$$

[11] Random variables are denoted by upper case letters, corresponding non-random variables by the same letter in lower case: for example, "the random variable *X* has probability density function *f(x)*".

When $n_s = N$, sampling without replacement is exhaustive and the variance is 0 (the sample is always the population and its mean is always precisely the population mean).

Appendix C: Fitting models to data

For all experiments we fit each individual participant's separately using maximum likelihood estimates. We used **fminsearchbnd** (J. D'Errico), a function based on **fminsearch** in MATLAB® (MathWorks) to search for the parameter values that minimized negative log likelihood, subject to limits described below. We estimated the parameter values maximizing likelihood by repeating the search process 1000–2000 times using different, randomly selected starting points and selected the parameter value estimates corresponding to the smallest value in negative log likelihood obtained. See Supplemental Table S1 for a summary of notations.

Experiment 1 (JRF)

1. **Bounded Log-Odds (BLO) Model**. The implementation of the BLO Model in JRF has seven free parameters, including four key parameters (bounds Δ^- and Δ^+, scaling parameter τ, anchor Λ_0) and three task-specific parameters (encoding-variance-related parameters n_s and κ, and noise variance σ_λ^2). Their limits were set as described below.

$\Delta^-, \Delta^+ \in \left[\ \lambda(0.01),\ \lambda(0.99)\ \right]$. Because the relative frequencies of the stimuli were in the range [0.01, 0.99], a Δ^- less than $\lambda(0.01)$ or a Δ^+ greater than $\lambda(0.99)$ would lead to the same log likelihood as a Δ^- equal to $\lambda(0.01)$ or a Δ^+ equal to $\lambda(0.99)$. Among the 11 participants, one participant's Δ^- and Δ^+ respectively reached the lower and upper limits.

$\tau \in (0, \infty)$. The scaling parameter.

Λ_0: In practice, we parameterized the product of τ and the anchoring parameter Λ_0 as the parameter corresponding to Λ_0 and required $\tau\Lambda_0 \in [-10,10]$. These limits were set for convenience. No participants reached the limits.

$n_s \in [1, N]$. Although n_s is described as sample size, we treated it as a continuous variable for convenience. Its upper limit was the number of samples in the stimulus. This limit varied from stimulus to stimulus over the range 200 to 600. None of the fitted values reached the upper limit, while 2 of the 11 participants hit the lower limit. We did not think it would be reasonable to judge relative frequency based on one sample, instead we considered the following possibility: When the n_s was estimated to be 1, the finite population correction factor in Eq. 8, $\frac{N-n_s}{N-1}$, would equal to 1, thus indistinguishable from a lack of finite population correction. That is, the participant compensated for their sampling variation as if the sampling was with replacement or from an infinite population, we would estimate $n_s = 1$.

$\kappa \in (0,\infty)$.

$\sigma_\lambda^2 \in (0,\infty)$.

2. **Bounded Prelec Model**. Its parameters and the limits on the parameters were the same as BLO, except that the limits for $[\Delta^-, \Delta^+]$ were $[\lambda'(0.01), \lambda'(0.99)]$ where $\lambda'(p) = -\log(-\log p)$ (Eq. 14), the counterpart of log-odds for Prelec's function.

3. **Bounded Linear Model**. Its parameters and the limits on the parameters were the same as BLO, except that the limits for $[\Delta^-, \Delta^+]$ were [0.01, 0.99], and the limits for Λ_0 were [0, 1], since all these parameters were now on the probability scale.

We tested a total of 18 models whose parameters were subsets of Δ^-, Δ^+, τ, Λ_0, n_s, κ, σ_λ^2 and whose limits can be derived from the limits specified above. See main text.

Experiment 2 (JRF)

The parameters and parameter limits for Experiment 2 were identical to those of Experiment 1, except that the limits of Δ^-, Δ^+ and n_s were adjusted to match the stimuli in Experiment 2—the range of relative frequency was [1/54, 53/54] and the maximum numerosity was 54. For BLO, the fitted parameters values of one of the 12 participants reached the upper limit of Δ^+ and no other limits were reached.

Experiment 3 (DMR)

The parameter settings for Experiment 3 were the same as those of Experiment 1 except that the sampling-related parameter n_s was omitted and there was an additional parameter α for the utility function, whose limits were set to be $(0,\infty)$. For BLO, among the 10 participants, the parameter fits for two participants reached the lower limit for Δ^-, while one participant's fit reached the lower limit of Λ_0, and no other limits were reached.

Experiment 4 (JRF and DMR)

The parameter settings for the JRF and DMR tasks of Experiment 4 were respectively identical to those of Experiments 1 and 3. For Experiment 4A that had two sessions, we fit the two sessions separately and together, as required by the analyses in the text.

Appendix D: Definitions for the additional models

1. CPT model with Tversky & Kahneman's (1992) weighting function. The weighting function has the form:

$$\pi(p) = \frac{p^{\gamma}}{\left(p^{\gamma} + (1-p)^{\gamma}\right)^{\frac{1}{\gamma}}}. \tag{A7}$$

where γ is a free parameter for the slope of probability distortion. The other assumptions of the model are identical to those of the other CPT models. In total, the model has three free parameters: $\gamma, \alpha, \sigma_{CE}^2$.

2. Transfer of attention exchange model (TAX, see Birnbaum, 2005). Similar to CPT, the TAX model assumes that the weight of a branch in the computation of expected utility depends on the branch's probability and that the weights of all branches add up to 1. But different from CPT, the weight for a specific branch is also influenced by the existence of the other branches in the gamble: branches compete for attention, which results in a transfer of weight from branches with higher-valued outcomes to those with lower-valued outcomes. For an *n*-branch gamble $(x_1, p_1; x_2, p_2; ...; x_n, p_n)$, where $x_1 > x_2 > \cdots > x_n > 0$, the probability weight for the *i*-th branch can be written as:

$$\pi_i = \frac{t(p_i) + (\delta/(n+1))\sum_{j=1}^{i-1} t(p_j) - (\delta/(n+1))\sum_{j=i+1}^{n} t(p_j)}{\sum_{i=1}^{n} t(p_i)}, \tag{A8}$$

where $t(p) = p^{\gamma}$, δ and γ are free parameters. In the case of two-outcome gambles $(x_1, p; x_2, 1-p)$, Eq. A8 is reduced to:

$$\pi_1 = \frac{\left(1-\frac{\delta}{3}\right)p^{\gamma}}{\frac{\delta}{3}p^{\gamma}+(1-p)^{\gamma}}. \qquad \text{(A9)}$$
$$\pi_2 = 1-\pi_1$$

That is, for two-outcome gambles, TAX is mathematically equivalent to a CPT model whose weighting function is defined by Eq. A9. The assumptions of the model for utility function and noise variance are identical to those of the CPT models. In total, the model has four free parameters: $\gamma, \delta, \alpha, \sigma_{CE}^2$.

3. Haruvy's regression model (Erev et al., 2010). Haruvy's regression model treats the prediction of human choices as a regression problem. It was originally a logistic regression model designed for the forced-choice task between a two-outcome gamble and a sure payoff. For gamble $G=(x_1,p;x_2,1-p)$ and sure payoff y, it defines the tendency to prefer the gamble as

$$T(G)=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 y+\gamma_1 p+\gamma_2 EV+\gamma_3 I, \qquad \text{(A10)}$$

where $EV = px_1+(1-p)x_2$ is the expected value of the gamble, I is an indicator variable that equals 1 if $EV > y$ and 0 otherwise, $\beta_0,\beta_1,\beta_2,\beta_3,\gamma_1,\gamma_2,\gamma_3$ are free parameters. The probability of choosing the gamble is modeled as a logistic function of $T(G)$:

$$\Pr(G)=\frac{1}{1+e^{-T(G)}}. \qquad \text{(A11)}$$

For the DMR task in the current paper, certainty equivalent (CE) is measured for each gamble. To apply Haruvy's regression model, we consider the problem as choosing between a gamble and its CE. The definition of CE implies that $\Pr(G)=0.5$, which requires $T(G)=0$. Thus the goal is to fit the following linear regression model:

$$0=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 CE+\gamma_1 p+\gamma_2 EV+\gamma_3 I+Normal\left(0,\sigma_{CE}^2\right). \qquad \text{(A12)}$$

It is obvious that one of the parameters in Eq. A12 is redundant. We arbitrarily set $\beta_3 = -1$. In total, the model has seven free parameters: $\beta_0, \beta_1, \beta_2, \gamma_1, \gamma_2, \gamma_3, \sigma_{CE}^2$.

Tables

Table 1. *Results of factorial model comparison for Experiments 1–4*

		Summed exceedance probability				
Dimen -sion	**Model family**	**Expt 1 JRF**	**Expt 2 JRF**	**Expt 3 DMR**	**Expt 4 JRF**	**Expt 4 DMR**
D1	Log-odds	71.8%	85.1%	23.2%	100.0%	99.9%
	Prelec	4.6%	4.1%	72.1%	0.0%	0.1%
	Linear	23.6%	10.8%	4.6%	0.0%	0.0%
D2	Bounded	87.4%	95.2%	92.9%	100.0%	100.0%
	Bounds-free	12.6%	4.8%	7.1%	0.0%	0.0%
D3	V(*p*)	91.5%	94.8%	88.6%	100.0%	100.0%
	$V \propto N$	4.9%	3.2%	N/A	0.0%	N/A
	V = const	3.5%	1.9%	11.4%	0.0%	0.0%

Note. Summed exceedance probability: The sum of the exceedance probabilities—the probability for a model to be the best model according to the Bayesian model selection method (Stephan et al., 2009)—of the specific family of models.

Supplemental Tables

Supplemental Table S1. *Notations*

General	
p	A generic value on the probability scale
λ	A generic value on the log-odds scale
$\lambda(p)$	Log-odds of probability *p*. $\lambda(p) = \log(p/(1-p))$
$\pi(p)$	Distorted probability (also denoted as $w(p)$, the decision weight, in decision tasks)
Linear in Log-Odds (LLO) Model	
γ	Slope of the observed linear transformation of log odds
p_0	Crossover point; controlling the intercept of the linear transformation of log odds
Bounded Log-Odds (BLO) Model	
Δ^-, Δ^+	Controlling the bounds of the log-odds representation
τ	Scaling parameter
Λ_0	Anchor parameter
ω_p	Reliability parameter that may vary with p, inversely related to encoding variance
$V(\hat{p})$	Encoding variance of the estimate $\hat{p}$
κ	Anchor precision parameter
$\Gamma(\bullet)$	Bounding operation
Δ	Input half-range
Alternative Models	
$\lambda^I(p)$	Converting *p* to the Prelec scale
$\lambda^O(p)$	Converting *p* to the linear scale
Judgment of Relative Frequency (JRF)	
n_s	Sample size; controlling $V(\hat{p})$
σ_λ^2	Noise variance of judgment on the log-odds scale
p	Relative frequency of black or white dots in the array
Decision-Making under Risk (DMR)	
α	Exponent of the power utility function
σ_{CE}^2	Noise variance of certainty equivalent
p	Probability of a specific outcome in a lottery

Supplemental Table S2. *Parameters of BLO for Experiments 1–4*

	Key parameters				Variance-related		Noise	Value
Experiment	Δ^-	Δ^+	τ	Λ_0	n_s	κ	σ_λ^2 / σ_{CE}^2	α
1 JRF	-1.27 (0.22)	1.32 (0.79)	3.33	0.056 (0.51)	96.12	963.5	0.41	N/A
2 JRF	-1.66 (0.16)	1.82 (0.86)	2.46	0.13 (0.53)	10.88	77.30	0.35	N/A
3 DMR	-1.53 (0.18)	2.20 (0.90)	0.91	-0.43 (0.39)	N/A	5.41	425.1 [ξ]	0.51
4 JRF	-1.60 (0.17)	1.45 (0.81)	2.51	0.040 (0.51)	256.9	39.85	0.45	N/A
4 DMR	-1.68 (0.16)	1.57 (0.83)	1.40	-0.21 (0.45)	N/A	8.40	510.1 [ξ]	0.74

Note. Medians across participants are reported. Value inside the parentheses is the probability equivalent of the parameter (originally on the log-odds scale).

[ξ] The noise parameter for DMR is denoted σ_{CE}^2, which is defined on a different scale than the σ_λ^2 of JRF.

Supplemental Table S3. *Studies used in the meta-analysis on the slope of probability distortion in decision under risk*

Study	γ	**Distribution of *p* in** $(x_1, p; x_2, 1-p)$
Abdellaoui, Diecidue, and Öncüler (2011)	0.62 (W3A)	1/6, 2/6, 2/6, 3/6, 4/6, 5/6
Abdellaoui, L'Haridon, and Paraschiv (2011)	0.65 (W2)	0.05, 0.25, 0.5, 0.75, 0.95
Bruhin, Fehr-Duda, and Epper (2010)	0.42^{ξ} (W2)	0.05×4, 0.1×3, 0.25×3, 0.5×6, 0.75×3, 0.9×3, 0.95×3
Glöckner and Pachur (2012)	0.61 (W1) 0.67 (W2)	180 gambles from Rieskamp (2008); 40 gambles from Glöckner and Betsch (2008); 36 gambles from Holt and Laury (2002) and Gächter, Johnson, and Herrmann (2007)
Gonzalez and Wu (1999)	0.44 (W2)	0.01, 0.05, 0.1, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9, 0.95, 0.99
Harrison and Rutström (2009)	0.91 (W1)	0.13, 0.25, 0.37, 0.5, 0.62, 0.75, 0.87
Rieskamp (2008)	0.77 (W1)	0.01, 0.02, …, 0.99
Stott (2006)	0.96 (W1) 0.96 (W2) 1.00 (W3A) 0.94 (W3B)	0.1×34, 0.2×14, 0.3×20, 0.4×20, 0.5×21, 0.6×18, 0.7×12, 0.8×14, 0.9×27
Tanaka, Camerer, and Nguyen (2010)	0.74 (W3B)	0.1, 0.3, 0.7, 0.9
Tversky and Kahneman (1992)	0.61 (W1)	0.01×2, 0.05×3, 0.1×3, 0.25×3, 0.5×6, 0.75×3, 0.9×3, 0.95×3, 0.99×2
Vrecko and Langer (2013)	0.72 (W1)	Same as Zeisberger et al., 2012 (see below)
Zeisberger et al. (2012)	0.86 (W1)	0.2×6, 0.5×12, 0.8×6

Note. All studies were selected from Table A.3 of Fox and Poldrack (2014) using the criteria described in the text. The γ column corresponds to γ^{+} in Fox and Poldrack (2014), with the codes in the parentheses denoting the functional form of the probability weighting function. Following Fox and Poldrack (2014), "W1" refers to Tversky and Kahneman's (1992) form,

"W2" refers to LLO, "W3A" and "W3B" refer respectively to one- and two-parameter Prelec functions. In the column of distribution of *p*, "*q*×*f*" (e.g. 0.05×4) denotes that the probability *q* is presented for *f* times (relative, not absolute count) in the gambles of the study.

[ξ] The value of γ listed here (0.42) is the result of Bruhin et al.'s (2010) ZH03 experiment, which is slightly different from the pooled value in Fox and Poldrack' (2014) Table A.3 (0.38). We could only use the ZH03 experiment because Bruhin et al. (2010) did not provide the full gamble set for their other experiments that we would need for our analysis.

Figure Legends

Figure 1. Inverted-*S*-shaped probability distortion and LLO fits. **a**. Subjective probability (decision weight) versus objective probability in decision under risk from Tversky and Kahneman (1992). **b.** Estimated relative frequency of letters in written English versus actual relative frequency from Attneave (1953). **c.** Estimated probability of signal present versus objective probability in a signal detection task from Tanner, Swets, & Green (1956), c.f. Green and Swets (1966/1974). **d–f.** The same data as **a–c**, re-plotted on the log-odds scale. Circles denote data. Thick curves or lines denote LLO fits. Adapted from Zhang & Maloney (2012).

Figure 2. Encoding on the Thurstone scale. A selected range of $\left[\Delta^-, \Delta^+\right]$ is encoded on the Thurstone scale with limited resolution. The smaller the range, the smaller the encoding variance.

Figure 3. Numerosity effect: data versus BLO prediction. Left and right columns are respectively for Experiments 1 and 2. Top and bottom rows are respectively for $\hat{\gamma}$ and $\hat{p}_0$. Each thin line denotes data for one participant. The thick black line with dots denotes mean data across participants. The thick gray line with dots denotes the mean predicted by BLO and the gray shadow denotes its 95% confidence interval. BLO predictions agreed well with data, capturing the (significant) declining trend in the observed $\hat{\gamma}$ and the (marginally significant) rising trend in the observed $\hat{p}_0$.

Figure 4. Results of factorial model comparison. a. Experiment 1, JRF, 11 participants. **b.** Experiment 2, JRF, 12 participants. **c.** Experiment 3, DMR, 10 participants. **d.** Experiment 4, JRF, 75 participants. **e.** Experiment 4, DMR, 75 participants. The

summed ΔAICc across participants is plotted for each model and experiment (task). Lower values of ΔAICc are better. We compare models that differ on three dimensions ("factors"), D1, D2 and D3. The horizontal axis corresponds to D1, the scale of transformation. Solid and dotted lines denote the two alternatives on D2, bounded or bounds-free. Different colors code alternative models on D3, variance compensation. BLO outperformed the alternative models in all experiments except Experiment 3, where a bounded, variance-compensated Prelec model was slightly better. In most cases, each assumption of BLO (log-odds, bounded, and $V(p)$) also outperformed the alternative assumptions on its dimension (e.g. solid lines are lower than dotted lines, other things being the same).

Figure 5. Data versus model predictions for individual participants of Experiment 4. a. $\hat{\pi}(p)-p$ versus p for JRF, with $\hat{\pi}(p)$ averaged across different numerosities. Each panel is for one participant. Black dots denote data. The red curves are the predictions of BLO, the blue curves the predictions of LLO ("linear in log-odds"). The last panel is the mean across participants. We plot $\hat{\pi}(p)-p$ instead of $\hat{\pi}(p)$ to better visualize the small but patterned probability distortions. **b.** Normalized CE ("certainty equivalent") for DMR, averaged across different outcome pairs. The format conventions are the same as in Fig. 5a. The corresponding plots in Figs. 5a and 5b are for the same participant. The BLO predictions were closer to data than the LLO predictions.

Figure 6. Invariances in the estimated parameters of BLO. Left and right panels of **a–c** are respectively for the JRF and DMR tasks. Each circle is for one participant, with participants in different experiments coded by different colors. **a.** Symmetry of Δ^{-} and

Δ^+ around 0. **b.** Invariance of the Thurstone constant $\Psi \equiv \tau\Delta$. The black curve denotes $\Psi \equiv \tau\Delta = C$, where C is the median value across participants. (11/85 data points are outside the plot range of the right panel.) **c.** Correlation of estimated Ψ across tasks in Experiment 4, where participants completed both JRF and DMR tasks. (5/75 data points are outside the plot range.) The r_s on the plot refers to Spearman's correlation coefficient, which is robust to outliers.

Figure 7. Maximizing mutual information. a–e. Expected mutual information between stimuli and responses is plotted against $-\Delta^-$ and Δ^+ as contour map, separately for each experiment (task). Higher values are coded as more reddish and lower values as more bluish. Number on a contour is its expected mutual information (in bits). For all experiments, the observed median value of $\left(-\Delta^-, \Delta^+\right)$ (marked by the red circle) was close to maximizing the expected mutual information. See text.

Figure 8. Meta-analysis supporting BLO's prediction on the slope of probability distortion in decision under risk. The estimated slope of probability distortion ($\hat{\gamma}$) is plotted against the standard deviation of the objective log-odds ($\lambda(p)$) of the gamble set, where p denotes the probability for the higher outcome of a two-outcome gamble, $\left(x_1, p; x_2, 1-p\right)$. Each data point is for one published study. The red line denotes the regression line. See text for the study selection and Supplemental Table S3 for a full list of the studies. Consistent with the prediction of BLO, $\hat{\gamma}$ decreases with the standard deviation of $\lambda(p)$.

Legends of Supplemental Figures

Figure S1. Relative-frequency judgment: Time course of one trial. Left: Experiment 1, which was reported as Experiment 1 in Zhang & Maloney (2012). The dots (0.26° in diameter) in the display were scattered uniformly at random within a 17°×17° area at the center of the screen. **Right**: Experiment 2. The dots (0.32° in diameter) in each display were scattered uniformly at random within a virtual circle of diameter 5°. The centers of the dots in the two displays were randomly chosen, separated by at least 6°.

Figure S2. Numerosity effect: data versus BLO prediction for individual participants. a. $\hat{\gamma}$ of Experiment 1. **b.** $\hat{\gamma}$ of Experiment 2. **c.** $\hat{p}_0$ of Experiment 1. **d.** $\hat{p}_0$ of Experiment 2. Each panel is for one participant. The black dots denote data. The red lines and error bars denote the means and 95% confidence intervals predicted by the bounded log-odds model.

Figure S3. Probability distortion predicted by BLO and LLO. a–e plot the model-predicted $\pi(p)$ against p for each experiment (task), with left and right panels respectively for BLO and LLO predictions. Each colored curve is for one participant. For JRF tasks, the predicted subjective probability is averaged across different numerosities. The bounds in the BLO fits introduce discontinuities in derivative ("kinks") into the fit. A red arrow in the first panel marks one such discontinuity.

Figure S4. Goodness-of-fit of BLO versus additional alternative models for DMR. The references used for computing ΔAICc were the same as those of Figure 4. There were 10 participants in Experiment 3 and 75 participants in Experiment 4. **a.** Models in the framework of the original Prospect Theory (OPT). The goodness-of-fit of OPT models

were worse than their counterpart CPT (Cumulative Prospect Theory) models. **b.** Models of alternative form of probability weighting function or outside the framework of Prospect Theory. TK = Tversky & Kahneman's (1992) weighting function, in the CPT framework. TAX = Transfer of Attention Exchange model (Birnbaum, 2005). HR = Haruvy's regression model (Erev et al., 2010). For Experiments 3 and 4, these alternative models fit either much worse than LLO (HR), or comparable to LLO (TK, TAX), but still much worse than BLO.

Figure S5. Experiment 4: Results of factorial model comparison separately for participants of $\hat{\gamma} \leq 1$ and $\hat{\gamma} > 1$. a. The JRF task. **b.** The DMR task. The mean ΔAICc across participants is plotted for each model and task. The $\hat{\gamma}$ refers to the estimated slope parameter of the LLO model. The horizontal axis corresponds to D1. Solid and dotted lines code alternative models on D2. Different colors code alternative models on D3. No matter whether $\hat{\gamma} \leq 1$ or $\hat{\gamma} > 1$, the BLO model outperformed the alternative models.

Figure S6. Interferences between JRF and DMR suggested by Experiment 4A versus 4B. In Experiment 4, 75 participants performed both the JRF and DMR tasks, with trials of the two tasks randomly mixed (Experiment 4A) or blocked (Experiment 4B). Performance differences between Experiments 4A and 4B would suggest task effects. **a.** JRF: Mean $\hat{\pi}(p) - p$ as a function of p separately for the five numerosity conditions. Each curve is for one numerosity. The values of $\hat{\pi}(p) - p$ seem to vary systematically with numerosity in Experiment 4B (right) but not in Experiment 4A (left). **b.** JRF: $\hat{\gamma}$ and $\hat{p}_0$ as functions of numerosity. Each colored line is for one participant.

The black line denotes the mean across participants. For Experiment 4B (right), the numerosity effect in $\hat{\gamma}$ (significant) and $\hat{p}_0$ (marginally significant) were similar to those of Experiments 1 and 2, while no numerosity effects were found for Experiment 4A (left). **c.** DMR: Mean normalized CE as a function of p separately for participants of Experiment 4A (in green), participants of Experiment 4B who performed DMR first (in cyan) and those who performed JRF first (in blue). Shadings denote SE.

Figure S7. Correlation of the Thurstone constant Ψ across time. Based on the results of Experiment 4A, where participants completed two sessions on two different days. Left and right panels are respectively for JRF and DMR. Each circle is for one participant. (10/51 data points are outside the plot range of the right panel.) The r_s on the plot refers to Spearman's correlation coefficient, which is robust to outliers.

Figure 1

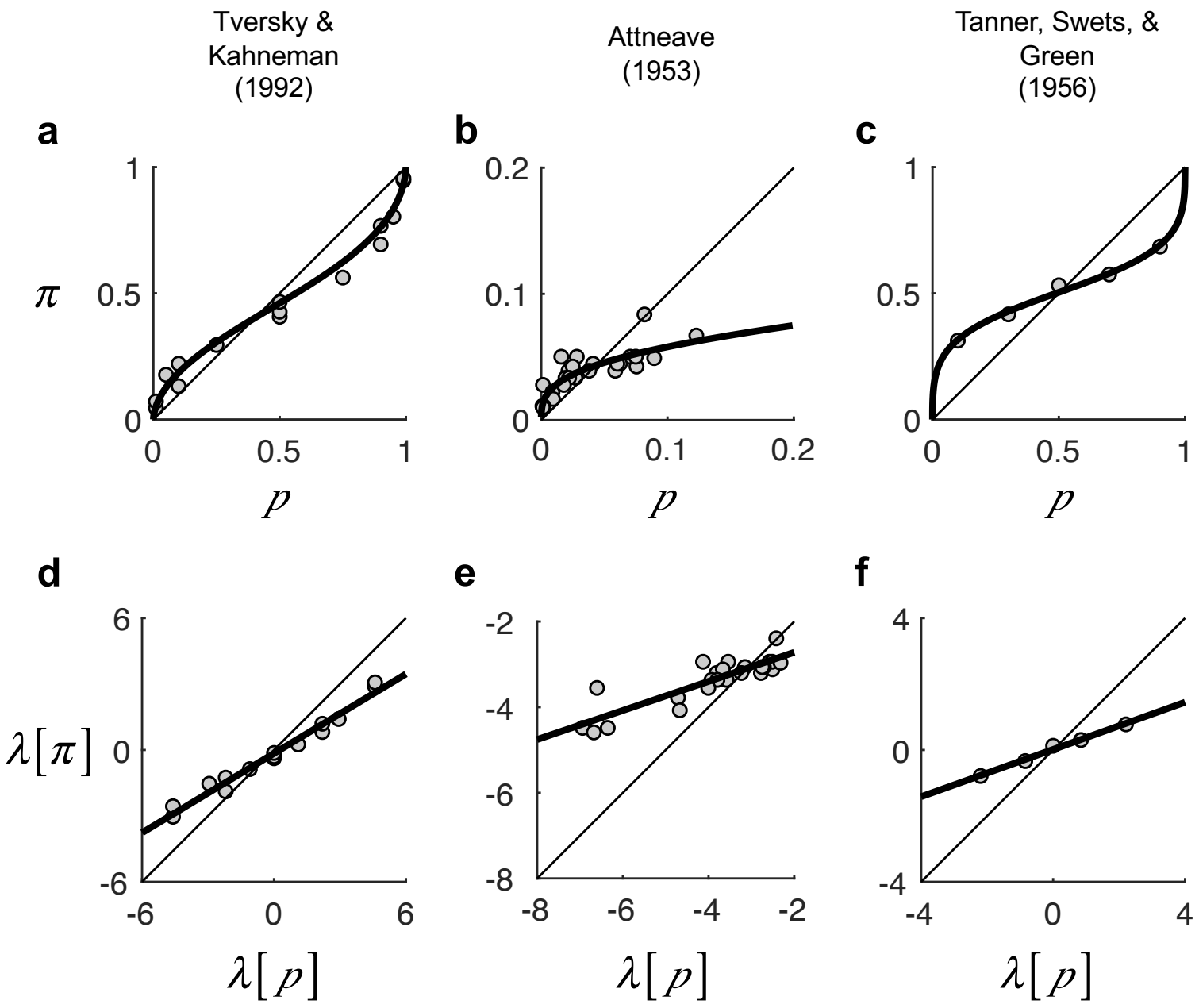

Tversky & Kahneman (1992)
Attneave (1953)
Tanner, Swets, & Green (1956)
a
b
c
d
e
f
π
p
λ[π]
λ[p]
1
0.5
0
0.2
0.1
6
-6
-2
-4
-8
4

Figure 2

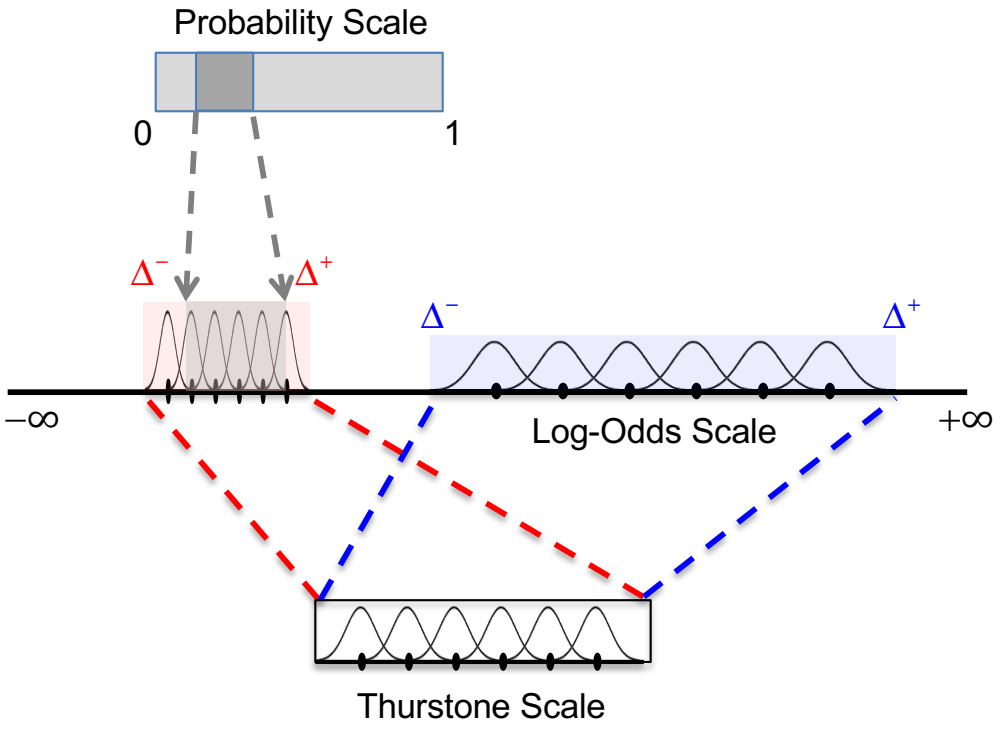

Probability Scale
0
1
Δ^-
Δ^+
Δ^-
Δ^+
$-\infty$
Log-Odds Scale
$+\infty$
Thurstone Scale

Figure 3

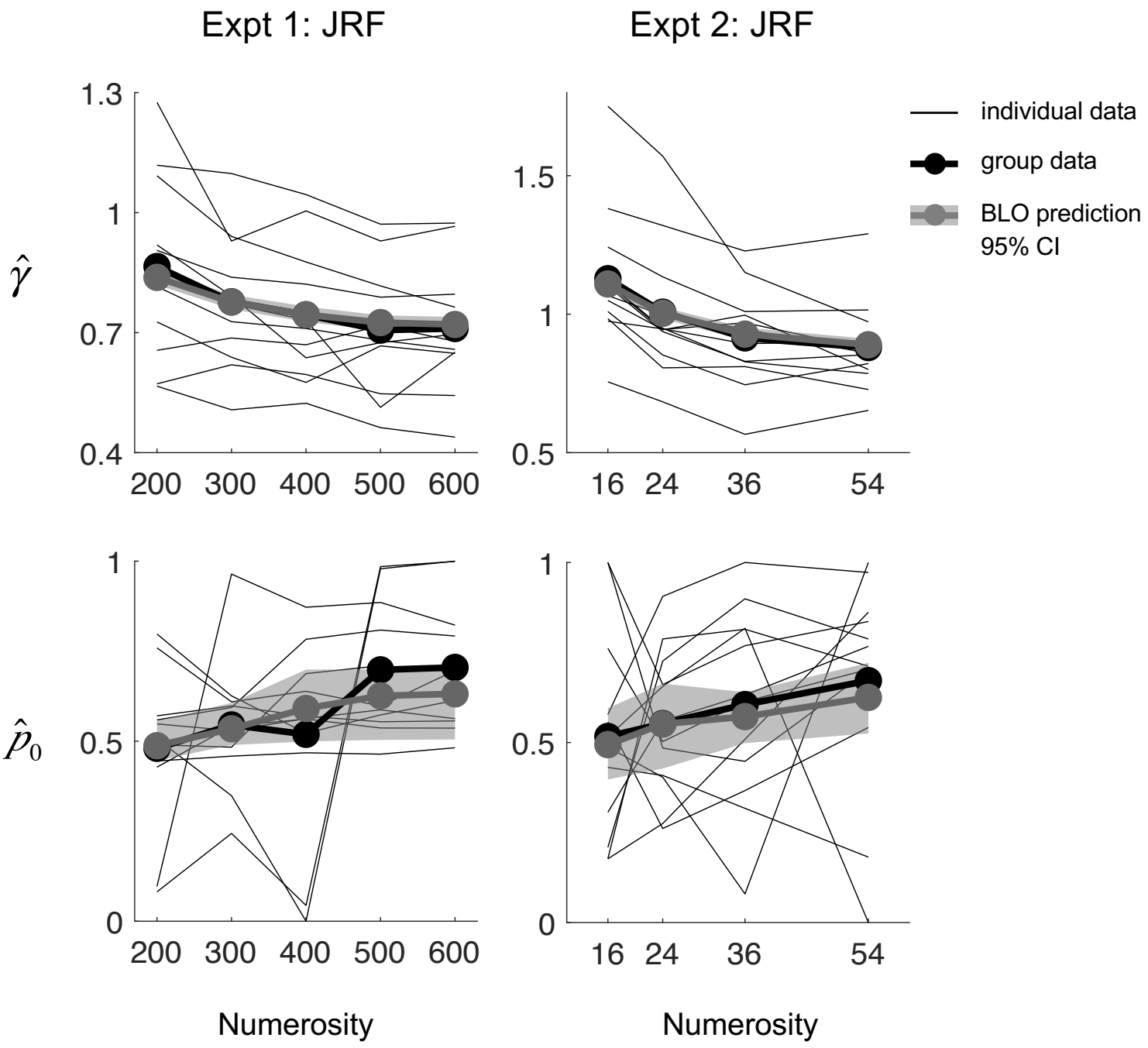

Expt 1: JRF
Expt 2: JRF
individual data
group data
BLO prediction
95% CI
$\hat{\gamma}$
$\hat{p}_0$
1.3
1
0.7
0.4
1.5
1
0.5
1
0.5
0
1
0.5
0
200 300 400 500 600
16 24 36 54
Numerosity
Numerosity

Figure 4

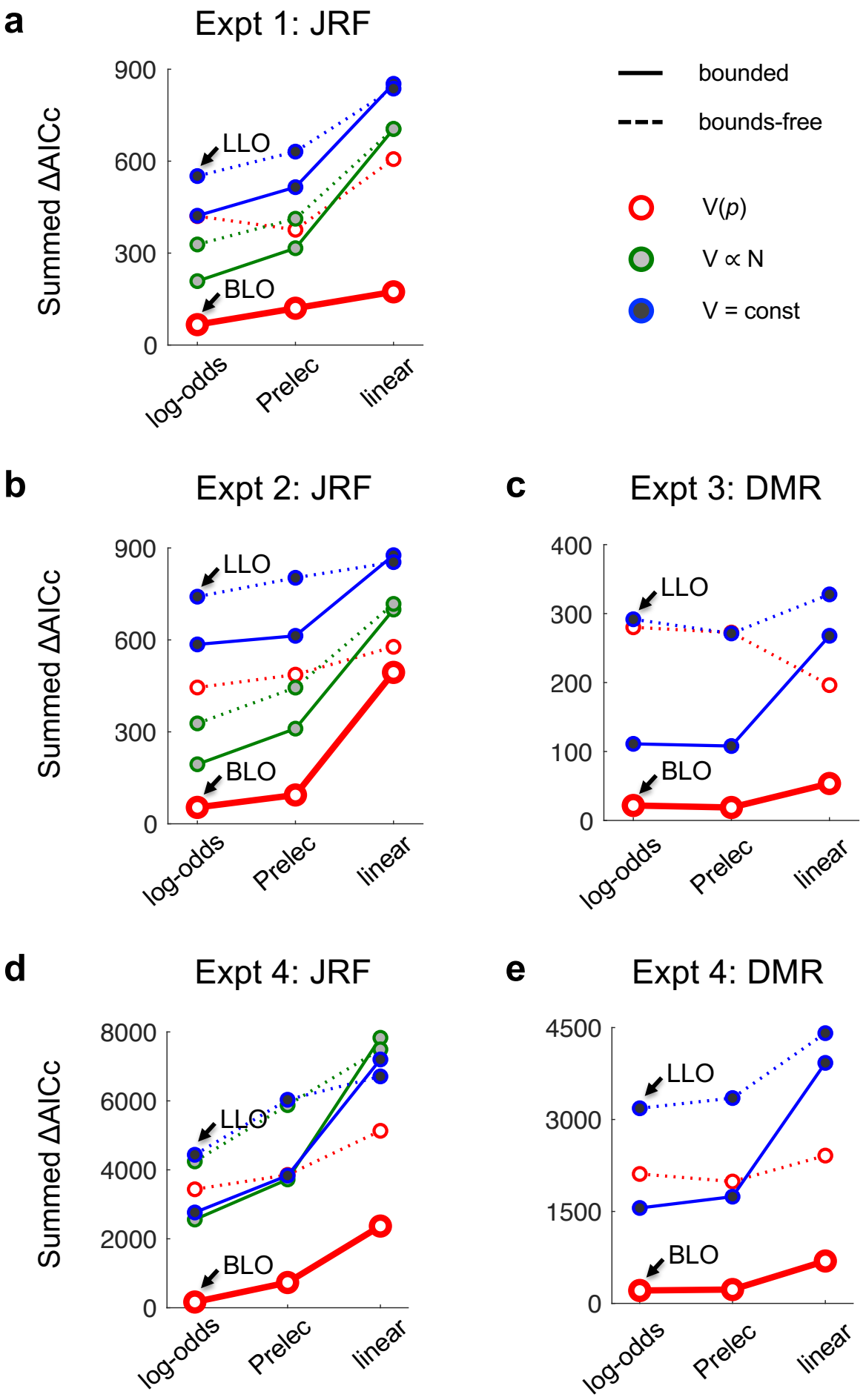

a
Expt 1: JRF
Summed ΔAICc
900
600
300
0
LLO
BLO
log-odds
Prelec
linear
bounded
bounds-free
V(p)
V ∝ N
V = const
b
Expt 2: JRF
Summed ΔAICc
900
600
300
0
LLO
BLO
log-odds
Prelec
linear
c
Expt 3: DMR
400
300
200
100
0
LLO
BLO
log-odds
Prelec
linear
d
Expt 4: JRF
Summed ΔAICc
8000
6000
4000
2000
0
LLO
BLO
log-odds
Prelec
linear
e
Expt 4: DMR
4500
3000
1500
0
LLO
BLO
log-odds
Prelec
linear

Figure 5

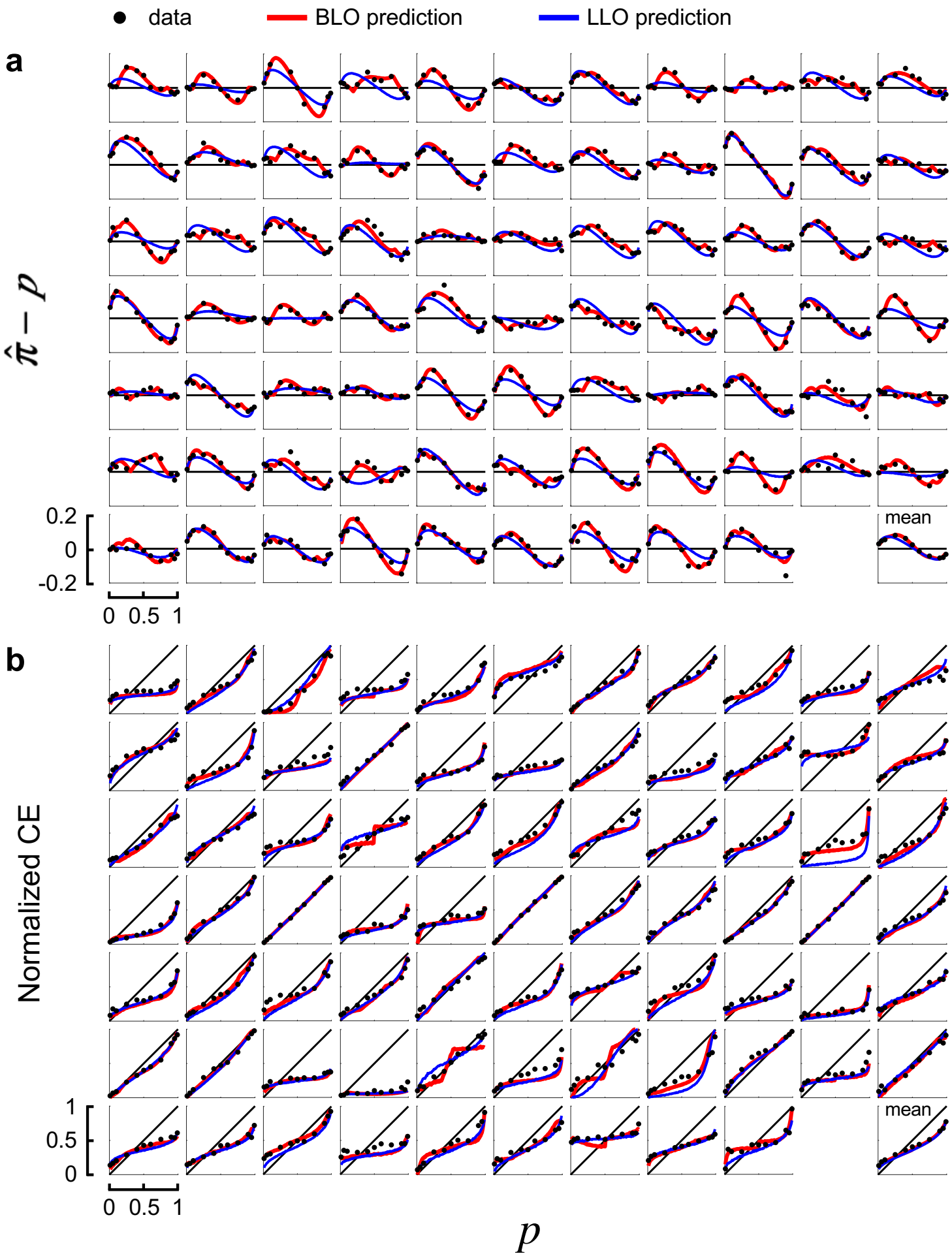

data
BLO prediction
LLO prediction
a
$\hat{\pi} - p$
0.2
0
-0.2
0
0.5
1
mean
b
Normalized CE
1
0.5
0
0
0.5
1
mean
p

Figure 6

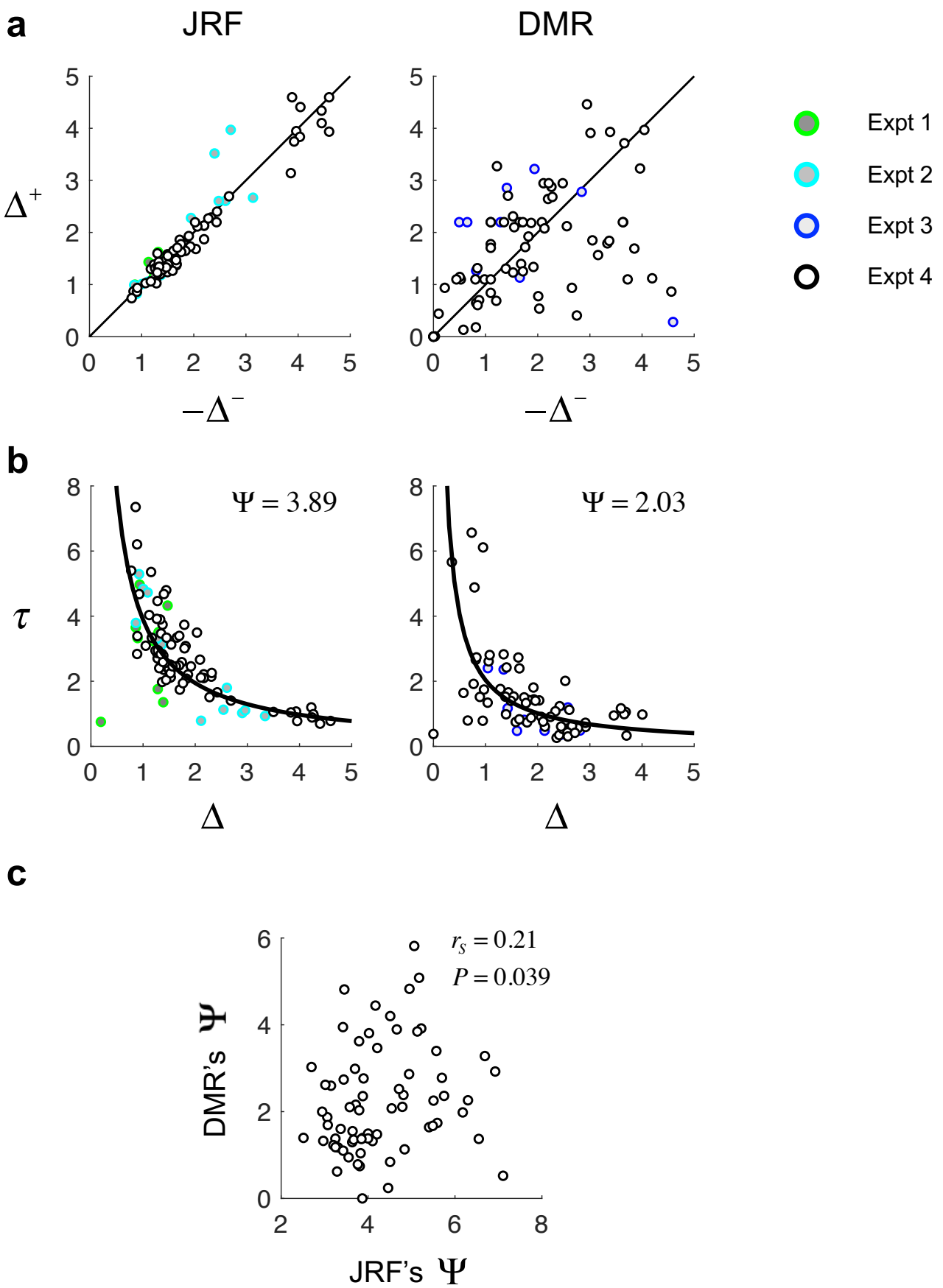

a
JRF
DMR
Δ^+
$-\Delta^-$
Expt 1
Expt 2
Expt 3
Expt 4
b
$\Psi = 3.89$
$\Psi = 2.03$
τ
Δ
c
$r_S = 0.21$
$P = 0.039$
DMR's Ψ
JRF's Ψ

Figure 7

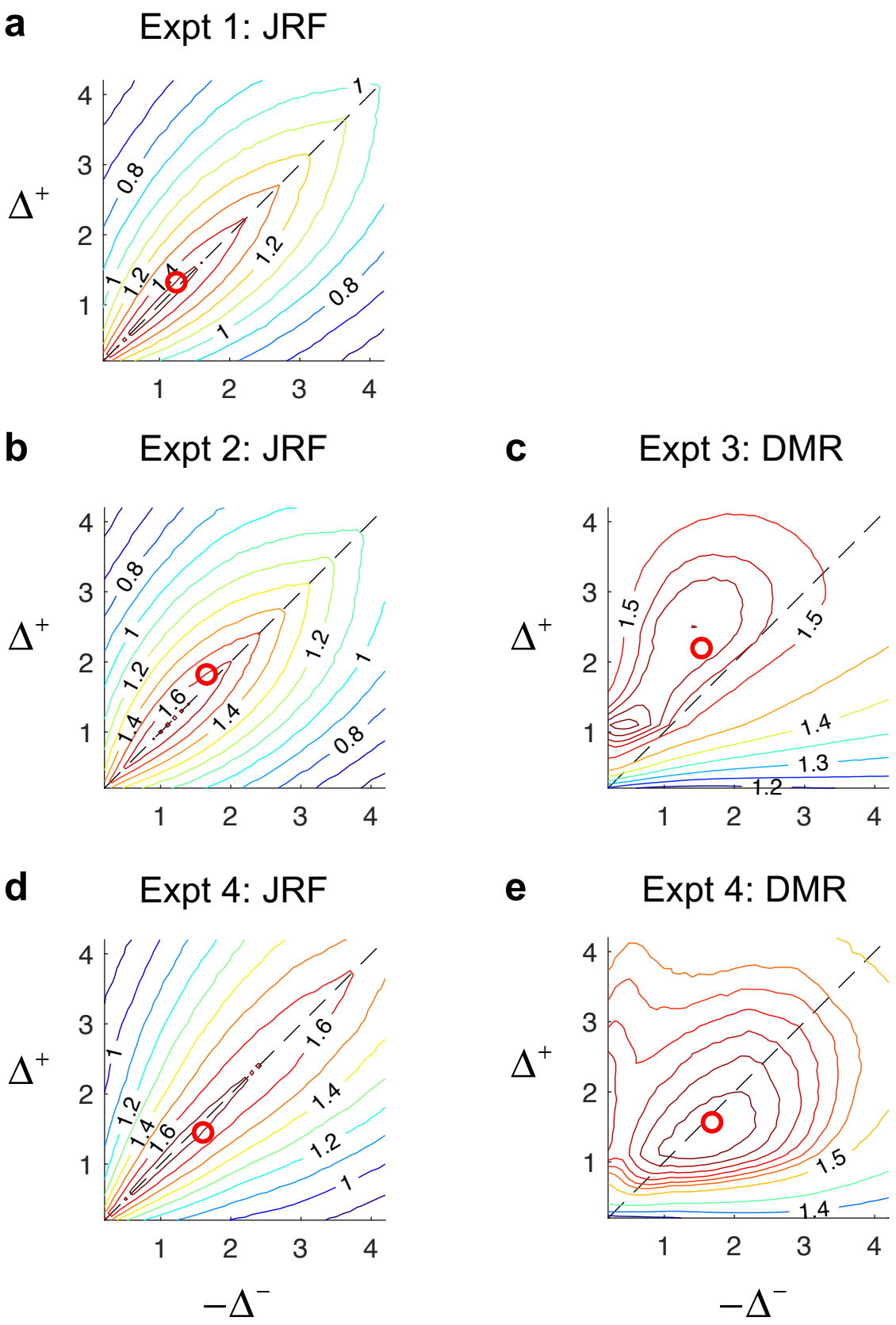
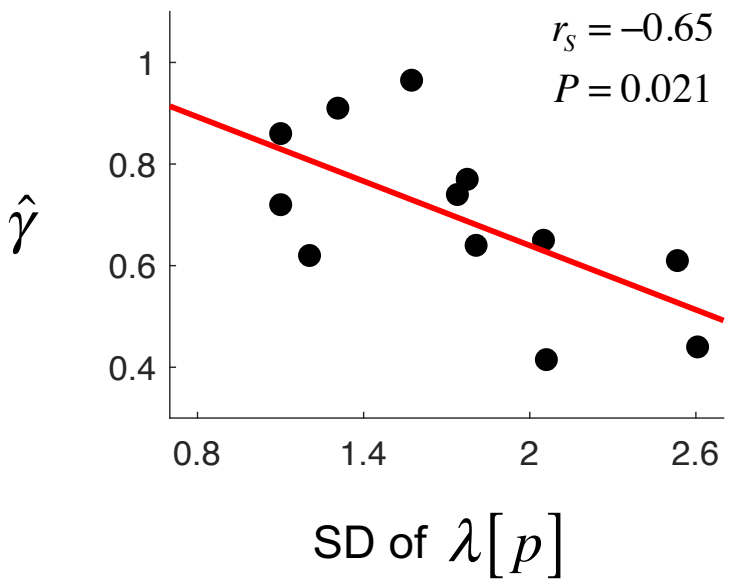
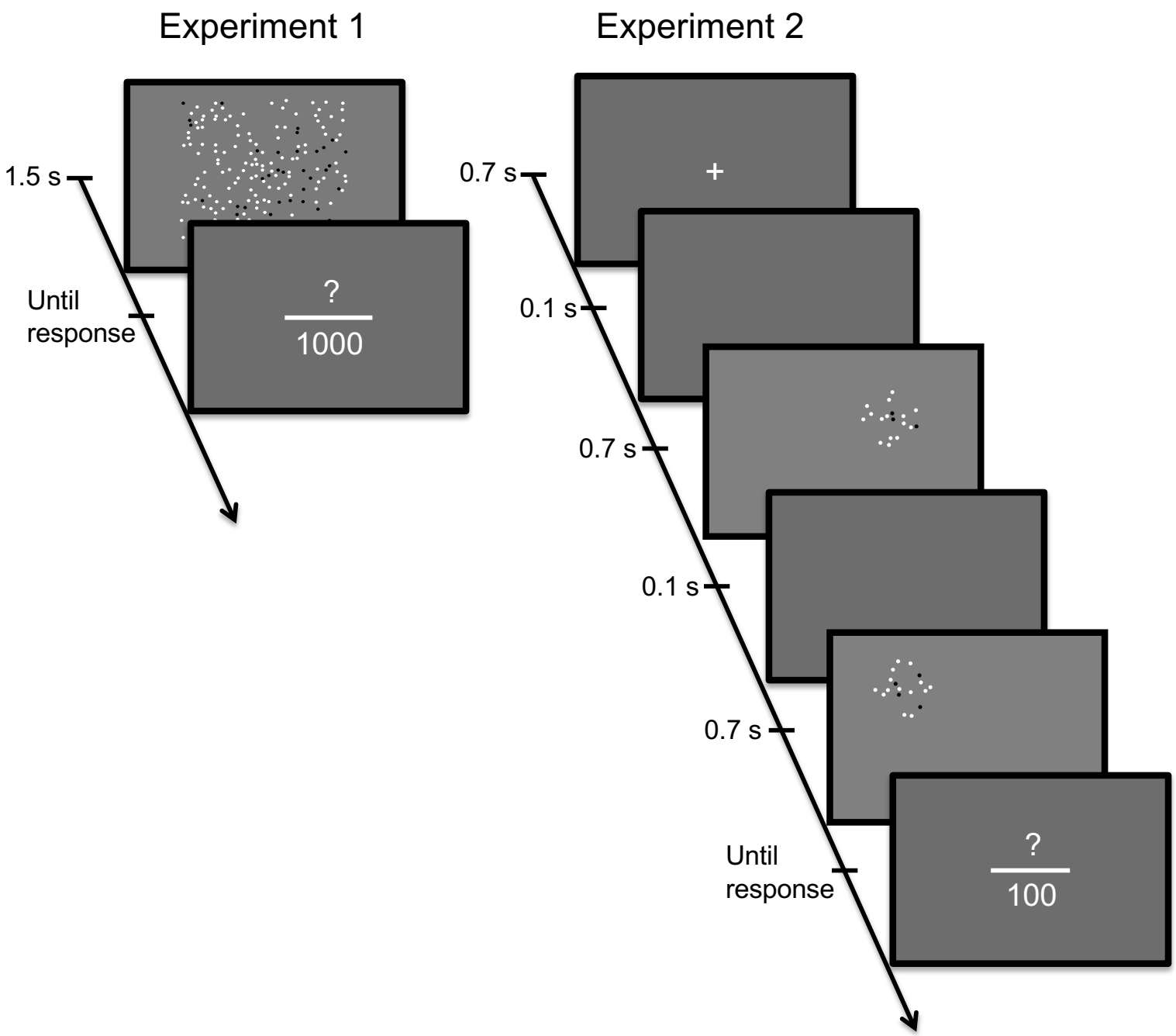

a
Expt 1: JRF
b
Expt 2: JRF
c
Expt 3: DMR
d
Expt 4: JRF
e
Expt 4: DMR
Δ^+
$-\Delta^-$

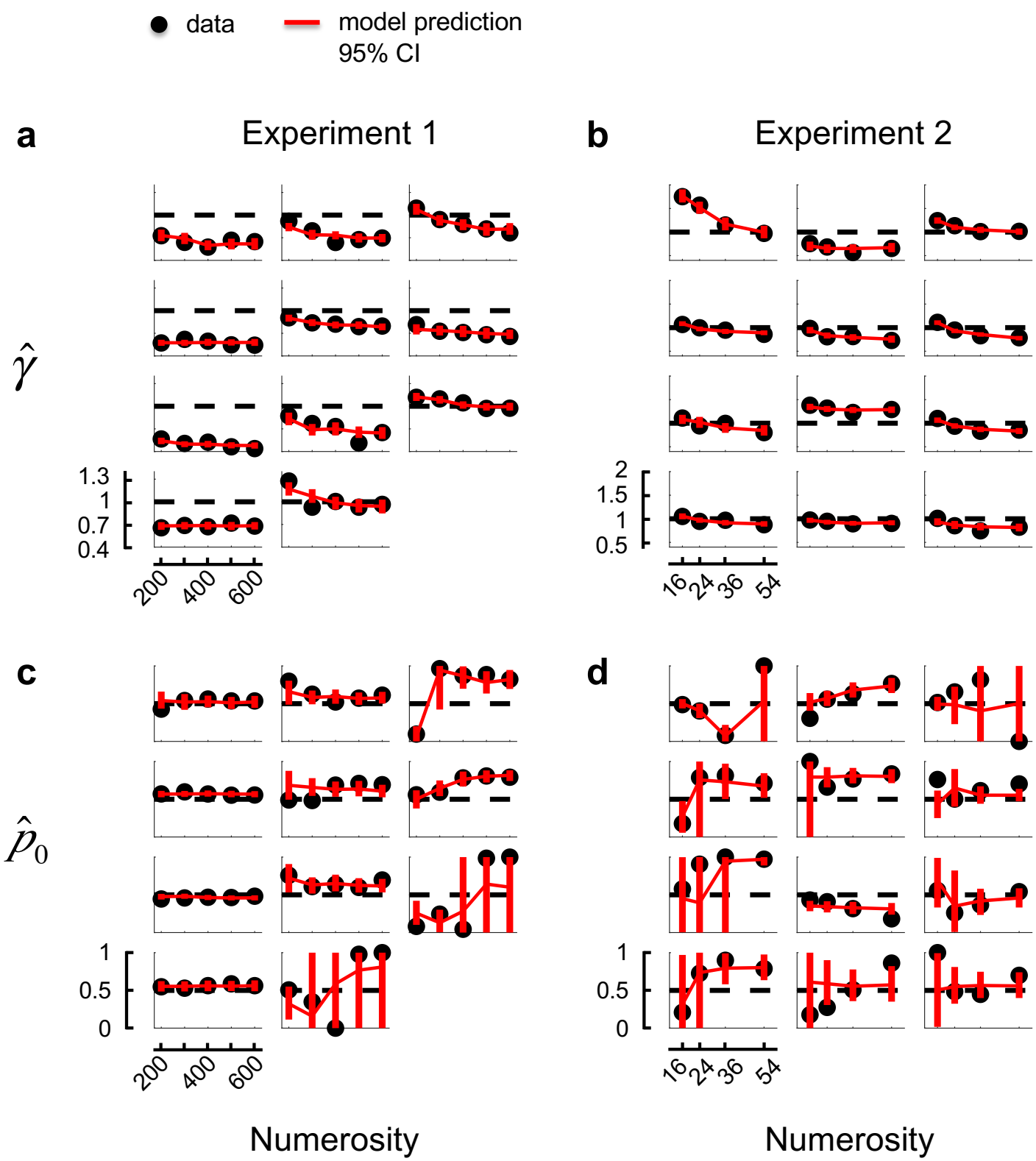
Figure 8



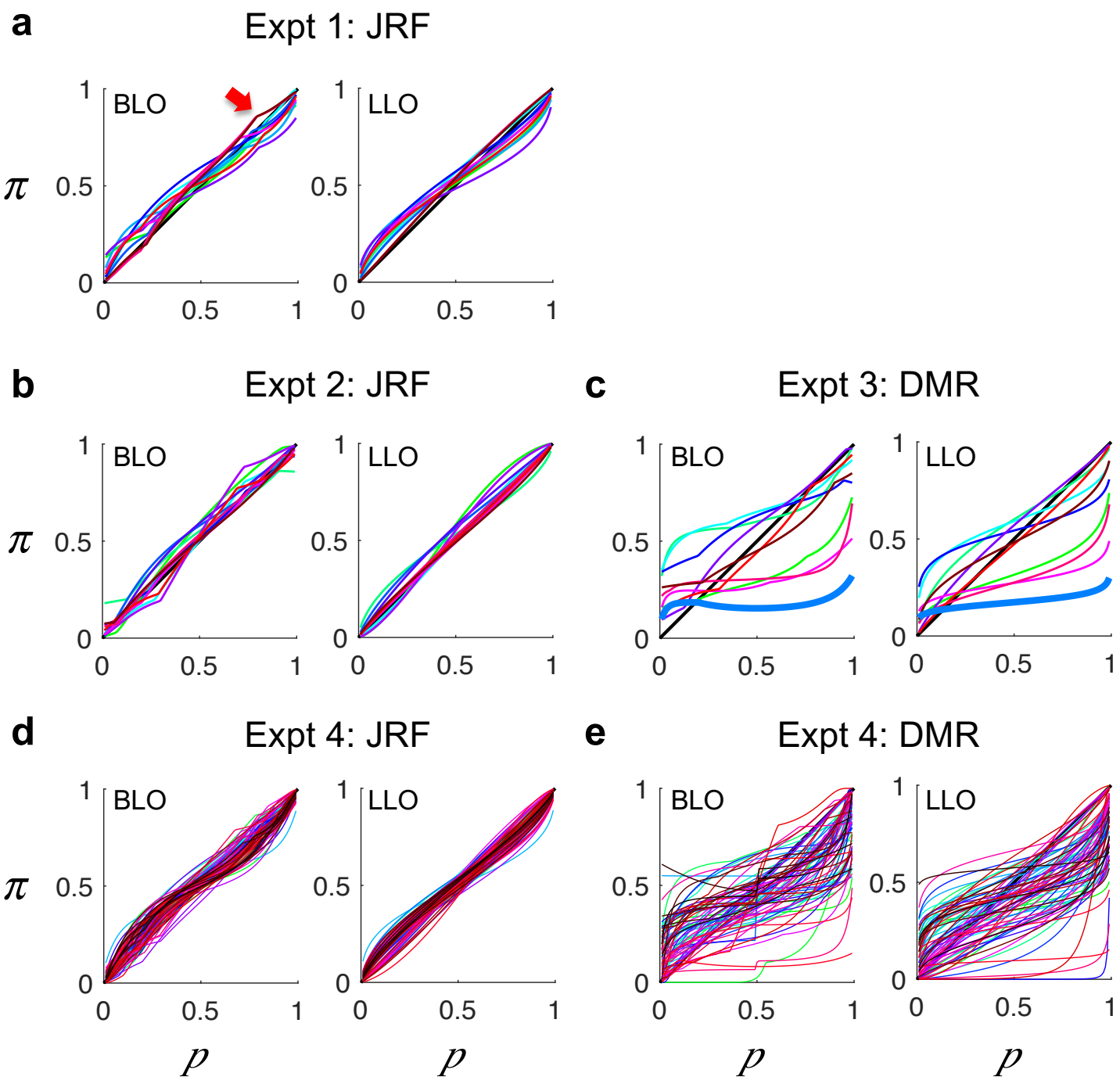
Supplemental Figure S1



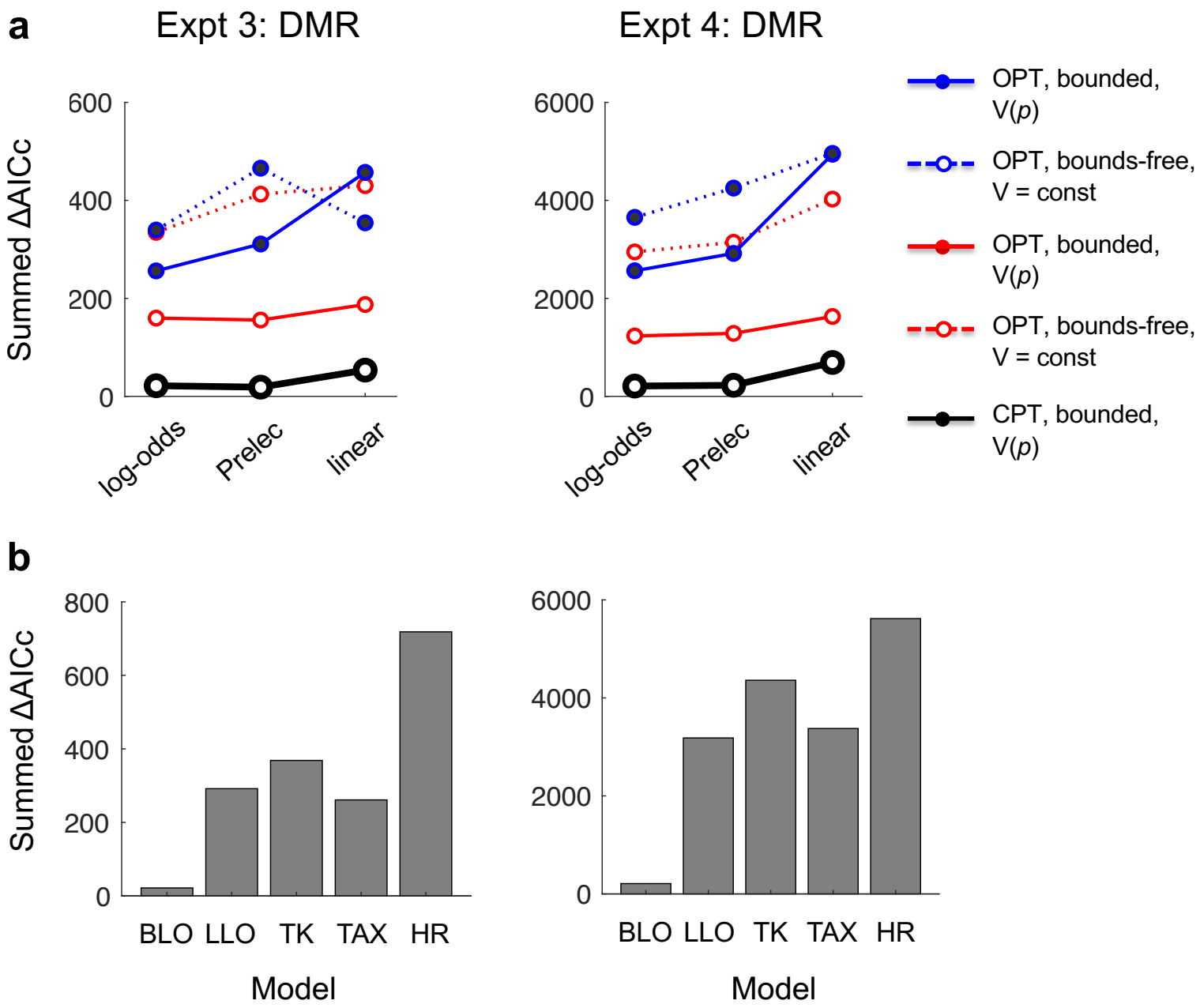
Supplemental Figure S2



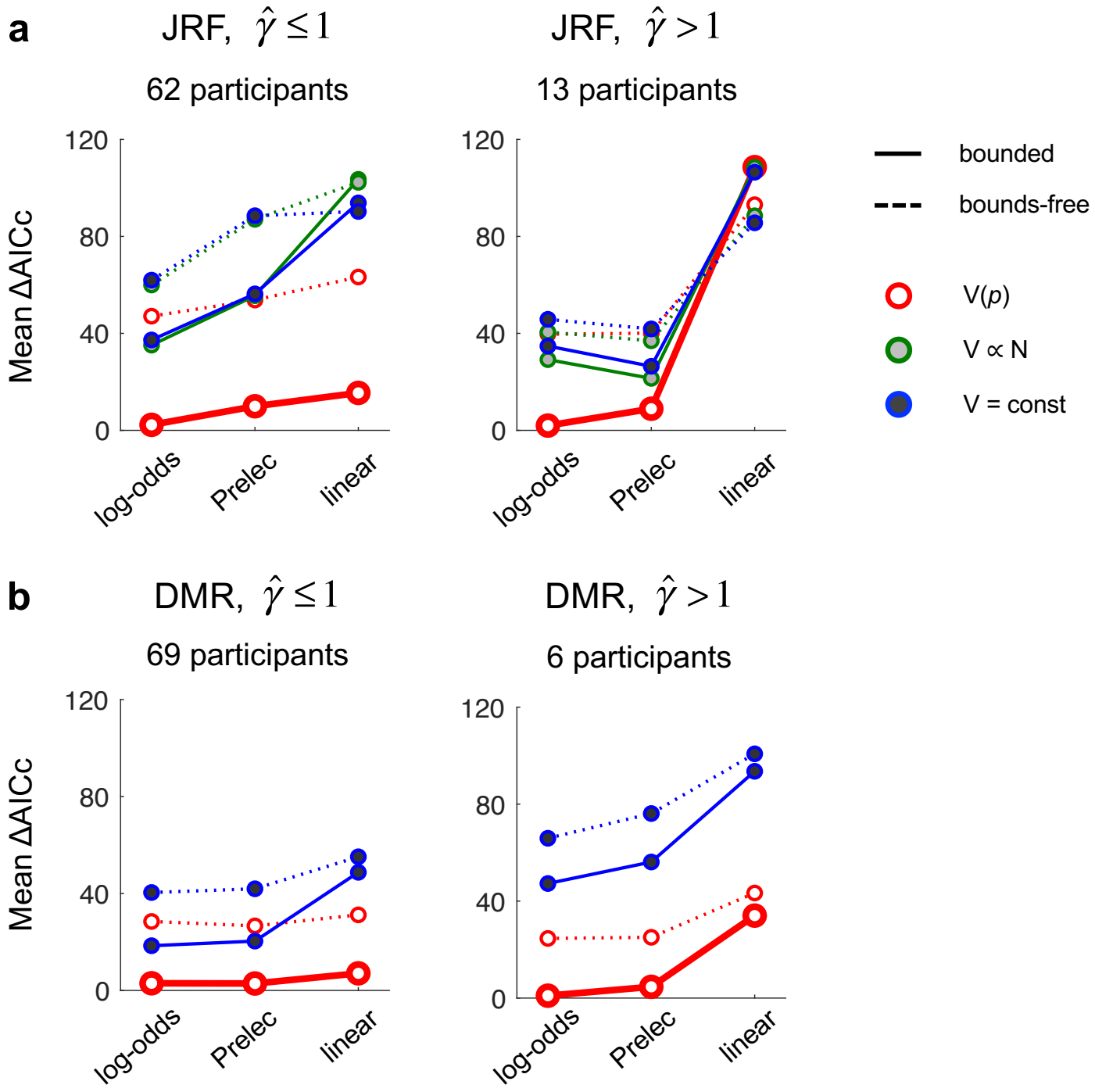
Supplemental Figure S3



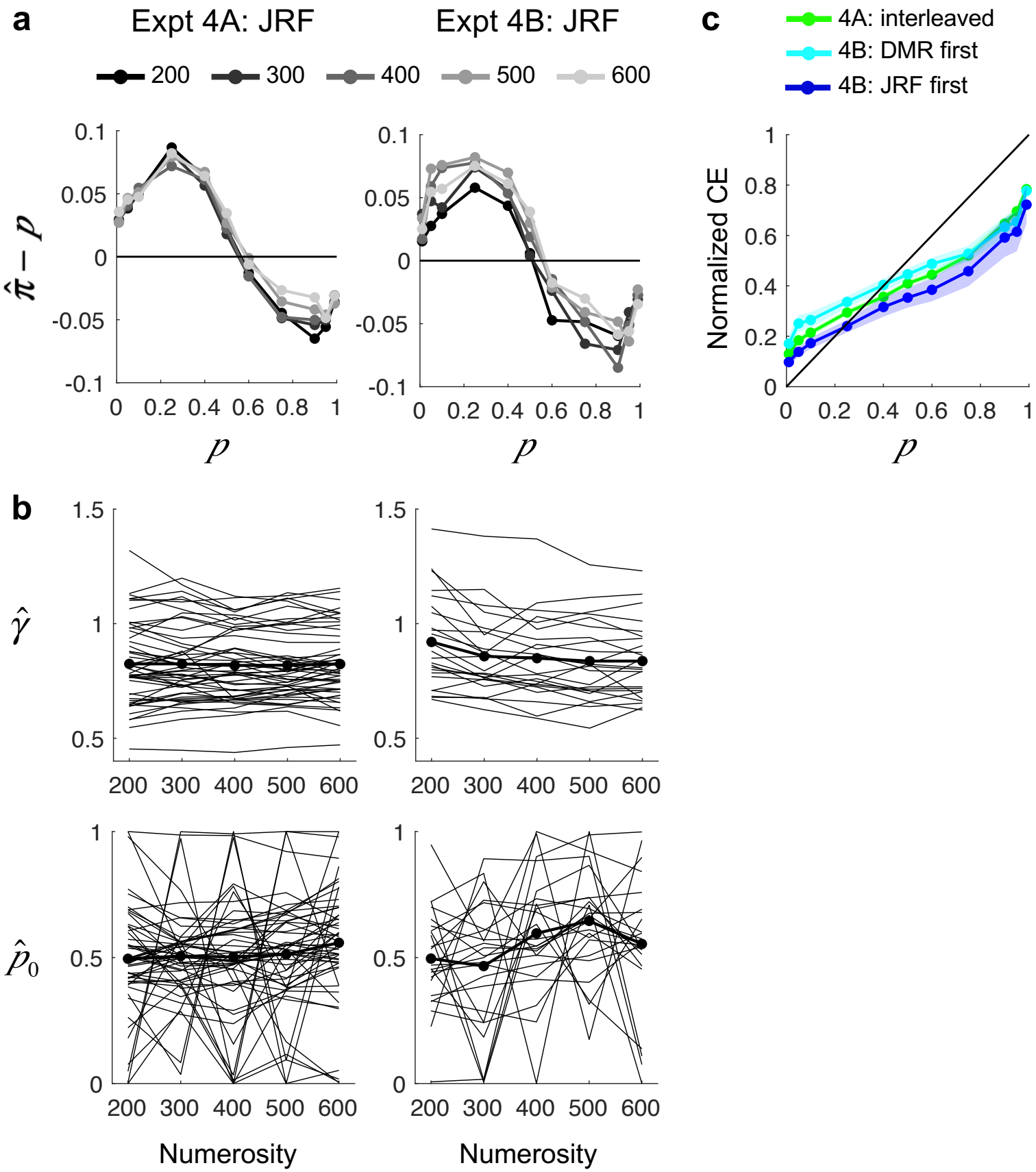
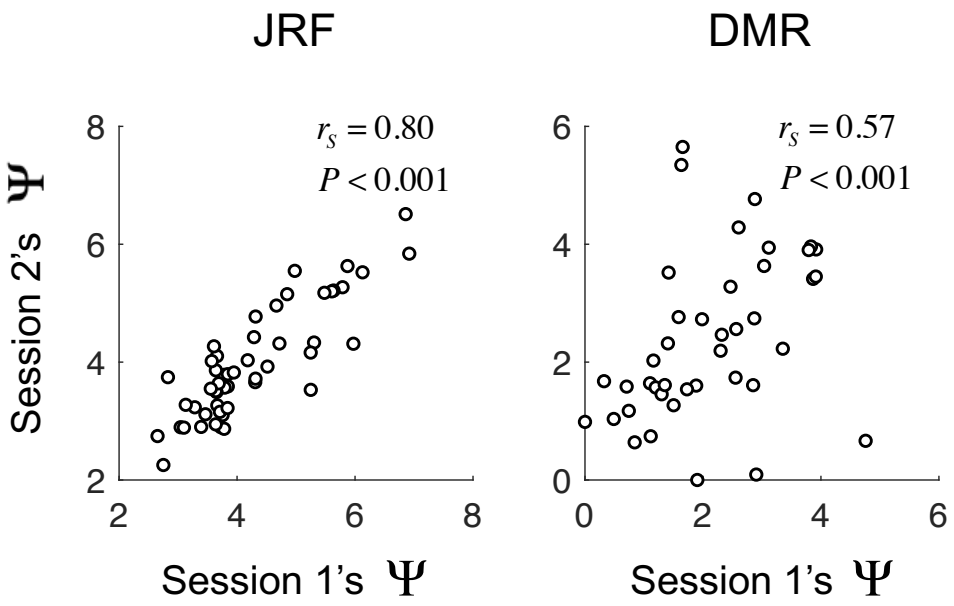
Supplemental Figure S4



Supplemental Figure S5



Supplemental Figure S6



Supplemental Figure S7