Automatic and fast encoding of representational uncertainty underlies probability distortion

Xiangjuan Ren[1,2] Huan Luo[3,4*] Hang Zhang[2,3,4*]

1. Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China

2. Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China

3. School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China

4. PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing, China

*Corresponding authors:

Huan Luo, email: huan.luo@pku.edu.cn

Hang Zhang, email: hang.zhang@pku.edu.cn

Keywords: judgment; decision making; probability distortion; MEG; steady-state response (SSR); time-resolved decoding

Short title: Encoding of representational uncertainty

Abstract

Humans do not have an accurate representation of the probability information in the environment but distort it in a surprisingly stereotyped way ("probability distortion"), as shown in a wide range of judgment and decision-making tasks. Many theories hypothesize that humans automatically compensate for the uncertainty inherent in probability information ("representational uncertainty") and probability distortion is a consequence of uncertainty compensation. Here we tested whether the representational uncertainty of probability is really quantified by the human brain and how its encoding contributes to probability distortion. Human subjects kept tracking the relative frequency of one color of dots in a sequence of dot arrays while their brain activity was recorded by magnetoencephalography (MEG). We found converging evidence from neural entrainment and time-resolved decoding that a mathematically-derived measure of representational uncertainty is automatically computed in the brain, despite it is not explicitly required by the task. In particular, the encodings of relative frequency and its representational uncertainty respectively occur approximately 300 ms and 400 ms after stimulus onset. Moreover, the relative strength of the brain responses to these two quantities predicts the subject's probability distortion behavior. The automatic and fast encoding of the representational uncertainty provides neural basis for the uncertainty compensation hypothesis of probability distortion. More generally, the representational uncertainty is closely related to confidence and our findings thus exemplify how confidence may emerge even earlier than and influence the final judgment.

Introduction

Humans do not have an accurate representation of the probability or relative-frequency information in the environment but distort it in a surprisingly stereotyped way. Typically, small probability is overestimated and large probability underestimated, which can be well fit by a linear-in-log-odds (LLO) model with two parameters (Zhang & Maloney, 2012). Such "probability distortion" phenomena occur in a variety of judgment and decision-making tasks such as relative-frequency estimation (Attneave, 1953; Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978; Varey, Mellers, & Birnbaum, 1990), confidence rating (Erev, Wallsten, & Budescu, 1994; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Wallsten, Budescu, Erev, & Diederich, 1997), decision under risk (Gonzalez & Wu, 1999; Luce, 2000; Tversky & Kahneman, 1992; Zhang, Ren, & Maloney, 2019), and are also widely reported in animal behaviors (Ferrari-Toniolo, Bujold, & Schultz, 2019; Stauffer, Lak, Bossaerts, & Schultz, 2015; Yang & Shadlen, 2007). However, the neural computations that cause such probability distortions remain largely unknown.

One computation that is assumed to be central to probability distortion (Fennell & Baddeley, 2012; Martins, 2006; See & Fox, 2006; Zhang et al., 2019) is compensation for the uncertainty inherent in the representation of noisy probability information ("representational uncertainty"). In particular, comparison among an extensive set of computational models of probability distortions (Zhang et al., 2019) suggests that the representational uncertainty in the form of $p(1–p)$ is compensated in human behaviors, where p denotes the probability or relative frequency to be represented. That is,

representational uncertainty is zero at p = 0 or 1 and maximal at p = 0.5. The $p(1–p)$ form had been reached by Lebreton et al. (2015) as a measure of uncertainty that correlates with the explicitly reported confidence for a specific judged value. As shown in their functional MRI experiments, during the valuation process, this uncertainty measure is automatically encoded in the ventromedial prefrontal cortex (vmPFC)—the brain region known for encoding value—even in the absence of explicit confidence rating. We thus hypothesize that the neural encoding of $p(1–p)$ is essential to the emergence of probability distortion. Here we ask whether such representational uncertainty of probability information is automatically encoded in the brain and whether this encoding is fast enough to precede the explicit judgment of probability and to influence probability distortion, as theories of probability distortion would expect (Fennell & Baddeley, 2012; Martins, 2006; See & Fox, 2006; Zhang et al., 2019).

In the present study, we designed a novel experimental paradigm where the representational uncertainty of probability information (i.e. $p(1–p)$) was varied continuously over time. By using time-resolved neural measurements and assessing the temporal coupling between stimulus variables and ongoing brain activities, we examined how the encoding of representational uncertainty proceeds in the brain and drives probability distortion. In particular, 22 human subjects were instructed to track the relative-frequency (p) of one color of dots in a sequence of dot arrays (Figure 1A), during which their brain activities were recorded by magnetoencephalography (MEG).

We first found that, though *p* was the only variable subjects needed to track and *p(1–p)* was task-irrelevant, the periodic changes of *p(1–p)* as well as *p* entrained neural rhythms, supporting an automatic encoding of representational uncertainty in the brain. By using a time-resolved decoding analysis to delineate the temporal course, we further found that the encoding of *p* and *p(1–p)* peaked respectively around 300 and 400 ms. Finally, the relative strength of the neural responses to the two variables (*p* and *p(1–p))* in the parietal region could predict the variation of probability distortion behavior across individuals. Our results thus provide neural evidence for an automatic, fast encoding of representational uncertainty in the brain and indicate how this encoding may influence the probability distortion observed in a wide range of human behaviors.

Results

As shown in Figure 1A, on each trial, subjects saw a sequence of displays consisting of isoluminant yellow and cyan dots, refreshed every 150 ms, and were instructed to track the relative-frequency of dots in one color (i.e. yellow or cyan). To ensure persistent tracking throughout the trial, the sequence of displays might end after some random duration and subjects needed to report the relative-frequency of the last display by clicking on a percentage scale.

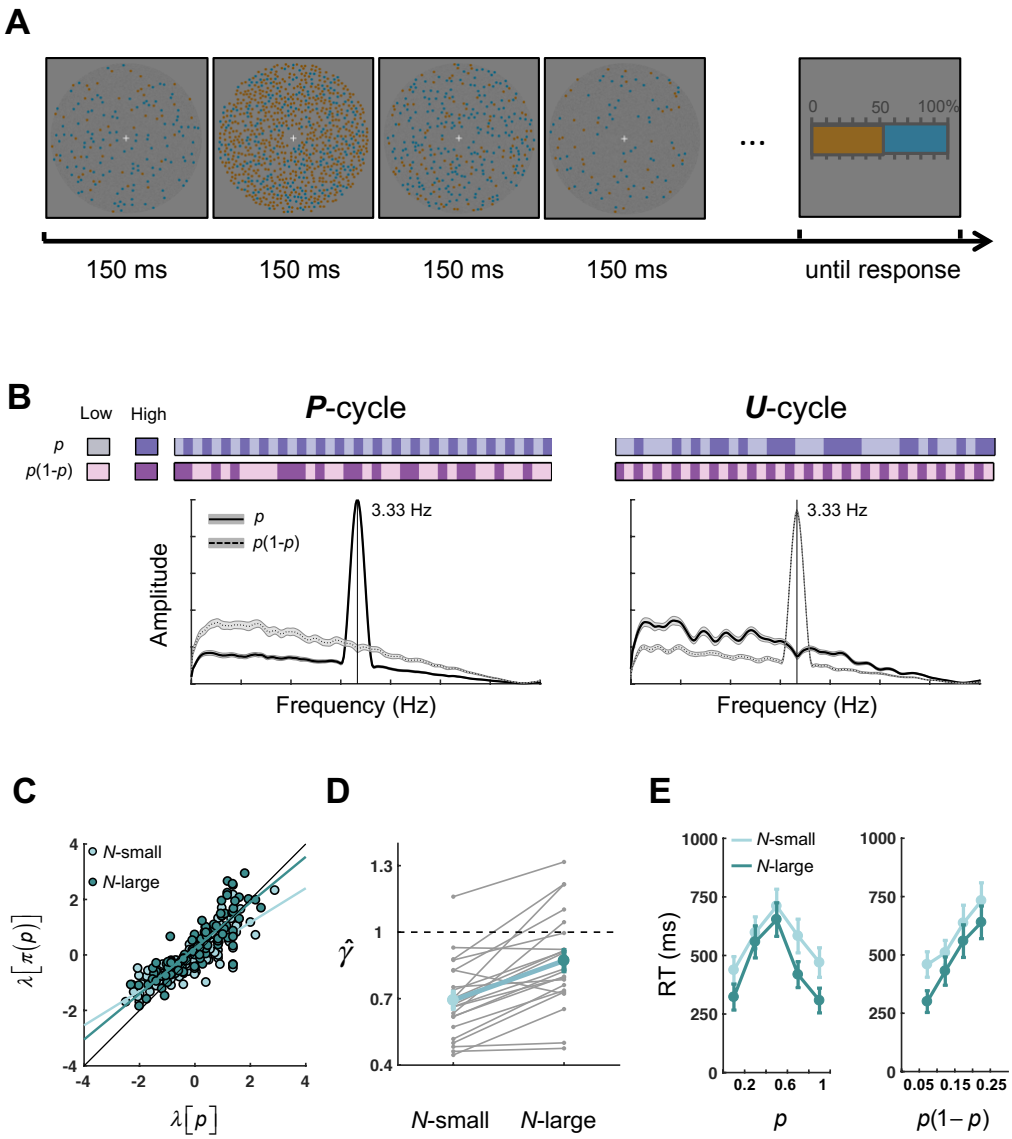

Figure 1. Experimental design and behavioral results. (A) Relative-frequency tracking and judgment task. On each trial, a sequence of yellow and cyan dot arrays was presented at a rate of 150 ms per display. Subjects were asked to fixate on the central fixation cross and track the relative frequency of yellow (or cyan) dots in each display until a 0–100% response scale appeared. Subjects then clicked on the scale to report the relative frequency in the last display. **(B)** The *p* and *p(1–p)* sequences in an example *P*-cycle or *U*-cycle trial. The value of *p* of each display was categorized into the smaller (<0.5) and the larger (>0.5), coded by different colors. Similarly, the smaller (<0.21) and the larger (>0.21) *p(1–p)* were coded by different colors. In *P*-cycles (left panels), the *p* sequence alternated between smaller and larger values, resulting in a frequency spectrum peaking at 3.33 Hz (solid line), while the *p(1–p)* sequence was aperiodic (dashed line). In *U*-cycles (right panels), it was the reverse: the *p(1–p)* sequence alternated between smaller and larger values in cycles of 3.33 Hz (dashed line), while the *p* sequence was aperiodic (solid line). **(C)** Distortion of relative frequency observed in one subject. The subject's reported relative-frequency $\pi(p)$ is plotted against the objective relative frequency *p*, both in the log-odds scale ($\lambda[p]=\log\frac{p}{1-p}$), separately for the *N*-small (light green) and *N*-large (dark green) conditions. Each dot denotes one trial. Lines denote linear fits (i.e. LLO model fits). **(D)** Estimated slope of distortion $\hat{\gamma}$ for different dot numerosity conditions. Each gray line is for one subject. The thick green line denotes the group mean, with error bars for SEM. The mean $\hat{\gamma}$ across the 22 subjects increased by 25.75% from the *N*-small to *N*-large condition. **(E)** Response time as an additional index for representational uncertainty. The mean response time (RT) for subjects to initiate their relative-frequency report (i.e. the interval between the response screen onset and the first mouse move) was plotted against binned *p* (left panel) or binned *p(1–p)* (right panel), separately for the *N*-small and *N*-large conditions. Error bars denote SEM. Consistent with our assumption that representational uncertainty is proportional to *p(1–p)*, the RT was an inverted *U*-shaped function of *p* and increased almost linearly with *p(1–p)*.

On each trial, the value of either *p* or *p(1–p)* was varied periodically at 3.33 Hz while the value of the other variable (i.e. *p(1–p)* or *p*) was controlled to be aperiodic (see Methods). The resulting two types of trials were referred as *P*-cycle and *U*-cycle

(*U* for uncertainty) trials respectively (Figure 1B). In other words, in *P-cycle* trials, the *p* value underwent a rhythmic fluctuation and the *p(1–p)* value followed an aperiodic random course, whereas the opposite pattern occurred for *U*-cycle trials. Note that subjects' task was always to attend to the relative frequency (the *p* value*)*, while the representational uncertainty of *p* (the *p(1-p)* value) was task-irrelevant. The values of *p* and *p(1-p)* were statistically independent of each other.

To dissociate the processing of the relative frequency and the total number of dots in a display (numerosity, denoted *N*), the numerosity was not constant but varied from display to display independently of *p* or *p(1–p)*, ranging from 10 to 90 (*N*-small trials) or from 100 to 900 (*N*-large trials). The introduction of *N*-small trials and *N*-large trials served as a means to manipulate the slope of probability distortion (Zhang & Maloney, 2012) so as to reveal the possible links between behavior and brain activity.

Behavioral and modeling results

All the 22 subjects performed well on reporting the relative-frequency of the last display, whose subjective estimate $\pi(p)$ was highly correlated with the objective relative-frequency *p* (all Pearson's $r > 0.71$, $P < 0.001$), suggesting that subjects had successfully tracked the change of *p* throughout the trial. As expected, most subjects showed the typical inverted-*S*-shaped "probability distortion", overestimating small *p* and underestimating large *p* (see Figure 1C for an example subject), a profile consistent with previous findings (Varey et al., 1990; Zhang & Maloney, 2012). We next quantified each subject's probability distortion using the two parameters

estimated from the LLO model (see Methods): slope $\hat{\gamma}$ and crossover point $\hat{p}_0$, separately for each cycle (*P*-cycle or *U*-cycle) and dot numerosity (*N*-small or *N*-large) condition. According to 2 (*P*-cycle, *U*-cycle) by 2 (*N*-small, *N*-large) repeated-measures ANOVAs separately performed for the estimated $\hat{\gamma}$ and $\hat{p}_0$, numerosity showed a significant influence on $\hat{\gamma}$ ($F(1, 63) = 57.82$, $P < 0.001$, $\eta_p^2 = 0.48$), with the $\hat{\gamma}$ estimated from *N*-large trials greater than that of *N*-small trials by 25.75% (Figure 1D), and a statistically significant but small influence on $\hat{p}_0$ ($F(1, 63)=4.24$, $P = 0.044$, $\eta_p^2=0.06$). No other main effects or interactions reached significance (all $P > 0.26$).

We found that the observed difference in $\hat{\gamma}$ between the *N*-small and *N*-large conditions could be well captured by the bounded log-odds model (Zhang et al., 2019) that assumes compensation for the $p(1-p)$ form of representational uncertainty (see Supplement S1 and Figure S1), with a greater $\hat{\gamma}$ implying lower representational uncertainty. That is, the *N*-large condition was associated with lower uncertainty than the *N*-small condition.

The response time (RT) of reporting relative frequency (defined as the interval between the response screen onset and the first mouse move) provides an additional index for representational uncertainty, given that lower uncertainty would lead to shorter RTs. We divided the values of *p* into five bins and computed RT for each bin and separately for the *N*-small and *N*-large trials. According to a 5 (*p* bins) by 2 (*N*-small vs. *N*-large) repeated-measures ANOVA on RTs, both the main effects of the *p* value ($F(4, 189) = 29.02$, $P < 0.001$, $\eta_p^2 = 0.38$) and numerosity ($F(1, 189) = 25.61$,

$P < 0.001$, η_p^2=0.12) were significant. Consistent with Lebreton et al. (2015), RTs were longest at p=0.5 and shortest when p was close to 0 or 1 (Figure 1E, left). More precisely, RTs increased almost linearly with $p(1–p)$ (Figure 1E, right), in agreement with what one would expect if representational uncertainty is proportional to $p(1–p)$. Besides, RTs were shorter in *N*-large trials than in *N*-small trials, which is in accord with the two conditions' difference in $\hat{\gamma}$ and provides further evidence that the *N*-large condition was accompanied by lower representational uncertainty.

Neural entrainment to periodic changes of *p* or *p(1–p)*

We next examined whether the brain response could track the periodically changing p values (*P*-cycle condition) or $p(1–p)$ values (*U*-cycle condition) in the stimulus sequence. Recall that the *P*-cycle and *U*-cycle trials had identical individual displays and differed only in the ordering of the displays: In *P*-cycle trials, p alternated between small and large values in cycles of 3.33 Hz while $p(1–p)$ was aperiodic; in *U*-cycle trials, $p(1–p)$ alternated in cycles of 3.33 Hz while p was aperiodic.

As summarized in Figure 2, the brain response indeed tracked the periodic changes of p and $p(1–p)$. In particular, in *P*-cycle trials (Figure 2, left), the phase coherence between the periodic p values and the MEG time series reached significance at 3.33 Hz (permutation test, *FDR* corrected $P_{FDR} < 0.05$), mainly for the occipital and parietal sensors. Surprisingly, significant phase coherence was also found at 3.33 Hz in *U*-cycle trials between the periodic $p(1–p)$ values and the MEG time series (Figure 2, right). That is, the brain activity was also entrained to the

periodic changes of the representational uncertainty of *p* (i.e. *p(1–p)*). Given that subjects were only asked to track the value of *p* but not *p(1–p)*, the observed neural entrainment to the task-irrelevant *p(1–p)* suggests an automatic encoding of representational uncertainty in the brain.

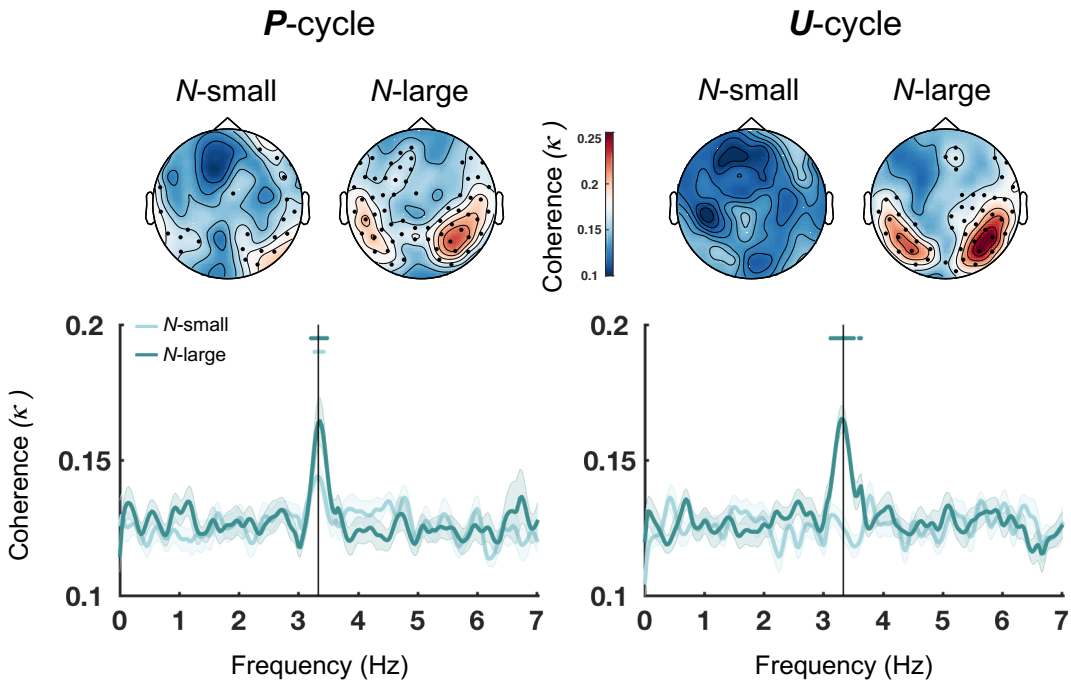

Figure 2. Results of phase coherence analysis. Grand-averaged phase coherence spectrum for magnetometers. Left: The phase coherence (κ) between the periodic *p* values and MEG time series in *P*-cycle trials. Right: The phase coherence between the periodic *p(1–p)* values and MEG time series in *U*-cycle trials. Light and dark green curves respectively denote the *N*-small and *N*-large conditions. Shadings denote SEM across subjects. The vertical line marks 3.33 Hz. Dots above the spectrum mark the frequency bins whose phase coherence was significantly above chance level (permutation tests, P_{FDR} <0.05 with *FDR* correction across frequency bins). **Insets:** Grand-averaged phase coherence topography at 3.33 Hz for each cycle and numerosity condition. Solid black dots denote sensors whose phase coherence at 3.33 Hz was significantly above chance level (permutation tests, P_{FDR}<0.05 with *FDR* corrected across magnetometers).

Notably, the significant phase coherences between the periodic *p* or *p(1–p)* values and MEG time series were not simply because the MEG time series consisted of 3.33 Hz frequency components. In a control analysis, we calculated the phase coherence between the aperiodic variable (i.e. *p* in *U*-cycle trials or *p(1–p)* in *P*-cycle trials) and the same MEG time series, for which there were no significant peaks at 3.33 Hz (Figure S5). Moreover, we had carefully controlled potential confounding variables in our experimental design. For example, *p* and *p(1–p)* were linearly independent of each other, both of which had negligible correlations with numerosity or low-level visual features such as luminance, contrast, and color variance (Figure S3A). We further confirmed that none of these variables led to different responses in *P*-cycles and *U*-cycles as reported above for *p* and *p(1–p)* (Figure S3B).

Behavioral probability distortion and the neural entrainment to *p* and *p(1–p)*

We next examined the relationship between behavioral probability distortion and the neural entrainment to *p* and *p(1–p)* on a subject-by-subject basis. Specifically, the phase coherence between a specific variable (*p* or *p(1–p)*) and the MEG time series can be considered as a measure of the strength of neural responses to the variable (Kramer & Eden, 2016). We defined

$$\beta = \frac{\kappa_p}{\kappa_{p(1-p)}} \tag{1}$$

to quantify the relative strength of *p* to *p(1–p)* in neural responses, where κ_p denotes the phase coherence for *p* in *P*-cycle trials, and $\kappa_{p(1-p)}$ denotes the phase coherence for *p(1-p)* in *U*-cycle trials. A higher value of β would imply a stronger

neural encoding of relative-frequency or a weaker encoding of representational uncertainty and is thus supposed to yield probability distortions of a greater slope (Zhang et al., 2019).

Given that the estimated slope of distortion $\hat{\gamma}$ was greater in *N*-large trials than in *N*-small trials, we would expect β to change in the same direction across the numerosity conditions. That is, suppose we define $\Delta\hat{\gamma} = \ln\left(\frac{\hat{\gamma}_{N\text{-large}}}{\hat{\gamma}_{N\text{-small}}}\right)$ and $\Delta\beta = \ln\left(\frac{\beta_{N\text{-large}}}{\beta_{N\text{-small}}}\right)$, there should be a positive correlation between $\Delta\hat{\gamma}$ and $\Delta\beta$. As shown in Figure 3, there was indeed a significant correlation between behavioral and neural measures across subjects in the parietal region (Pearson's r = 0.67, *FDR*-corrected one-tailed P_{FDR} = 0.033), associating the neural entrainment to *p* and *p(1–p)* with subsequent probability distortion behaviors.

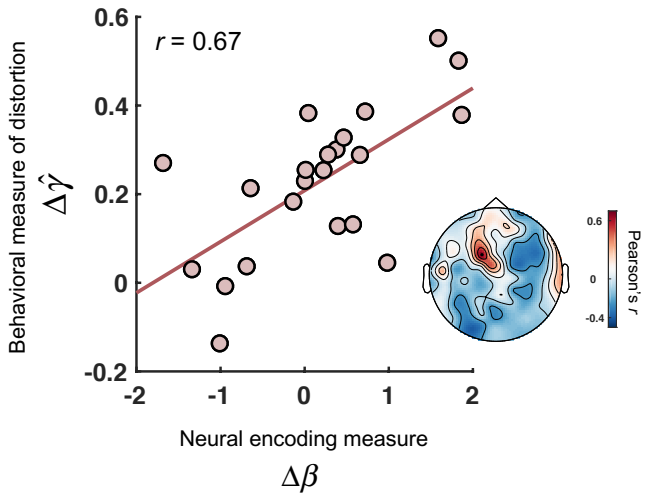

Figure 3. Neural responses to *p* and *p(1–p)* predict the slope of probability distortion. We defined $\Delta\hat{\gamma} = \ln\left(\frac{\hat{\gamma}_{N\text{-large}}}{\hat{\gamma}_{N\text{-small}}}\right)$ and $\Delta\beta = \ln\left(\frac{\beta_{N\text{-large}}}{\beta_{N\text{-small}}}\right)$ to respectively quantify how much the behavioral measure of the slope of distortion, $\hat{\gamma}$, and the relative strength

of *p* to *p(1–p)* in neural responses, β, changed across the two numerosity conditions.

Main plot: Correlation between $\Delta\hat{\gamma}$ and the $\Delta\beta$ at the parietal magnetometer channel MEG0421, Pearson's r = 0.67, P_{FDR} = 0.033. Each dot denotes one subject. **Inset**: Correlation coefficient topography between $\Delta\hat{\gamma}$ and $\Delta\beta$, on which MEG0421 is marked by a solid black dot.

Neural encoding of *p* and *p(1–p)* over time and the associated brain regions

Our experiment was designed such that on each trial one of the two variables of our interest (*p* or *p(1–p)*) changed periodically and the other was aperiodic. The encoding of a periodic variable led to periodic brain responses that were specifically phase-locked to the variable, as revealed by the phase coherence analysis above. However, it remained unknown how the aperiodic *p(1–p)* in *P*-cycles or aperiodic *p* in *U*-cycles was encoded over time.

To address this issue, we performed a time-resolved decoding analysis based on all 306 sensors (see Methods) using a regression approach that has been employed in a variety of EEG and MEG studies (Crosse, Liberto, Bednar, & Lalor, 2016; Wyart, Gardelle, Scholl, & Summerfield, 2012) including our previous studies (Huang, Jia, Han, & Luo, 2018; Jia, Liu, Fang, & Luo, 2017; Liu, Wang, Zhou, Ding, & Luo, 2017). The intuition of the time-resolved decoding analysis is as follows. Suppose the onset of each new *p* or *p(1–p)* value in the stimulus sequence evokes phase-locked brain responses that extend across time and whose magnitudes are proportional to the encoded value. The resulting MEG time series is a superposition of the responses to all the preceding stimuli in the sequence. But as soon as the values of an encoded

variable are not correlated over time (i.e. free of autocorrelation), their brain response profiles are separable. In particular, for any stimulus in the sequence, we can use the MEG signals at a specific delay after the stimulus onset to predict the value of the stimulus. Importantly, different from the phase coherence analysis (Figure 2) that only reflects the overall strength of the brain responses, this time-resolved decoding analysis allows us to assess how *p* and *p(1–p)* were encoded over time.

Figure 4 plots the temporal course of the decoding performance for *p* (left) and *p(1–p)* (right). Both *p* and *p(1–p)* were successfully decoded from the MEG signals in *N*-large trials (cluster-based permutation test, $P_{cluster} < 0.05$). In particular, the decoding performance for *p* peaked around 327 ms after stimulus onset (Figure 4, left), while that for *p(1–p)* peaked around 419 ms (Figure 4, right). In other words, the neural encodings of relative frequency and its representational uncertainty had distinct time courses, with the latter occurring approximately 100 ms later than the former. None of the potential confounding variables (numerosity, luminance, etc.) showed temporal courses similar to *p* or *p*(1-*p*) when the same time-resolved decoding procedure was applied (Figure S3C).

Based on the time windows that achieved the highest time-resolved decoding performance, we further performed a similar decoding analysis at each sensor location (including one magnetometer and two orthogonal planar gradiometers) separately for *p* and *p(1–p)* to assess the associated brain regions (see Methods). We found that the encoding of the two variables involved overlapping parietal-occipital regions (Figure 4, see Figure S6 for the results of the other above-chance time

windows), including the parietal region where $\Delta\beta$ and $\Delta\hat{\gamma}$ were positively correlated (Figure 3).

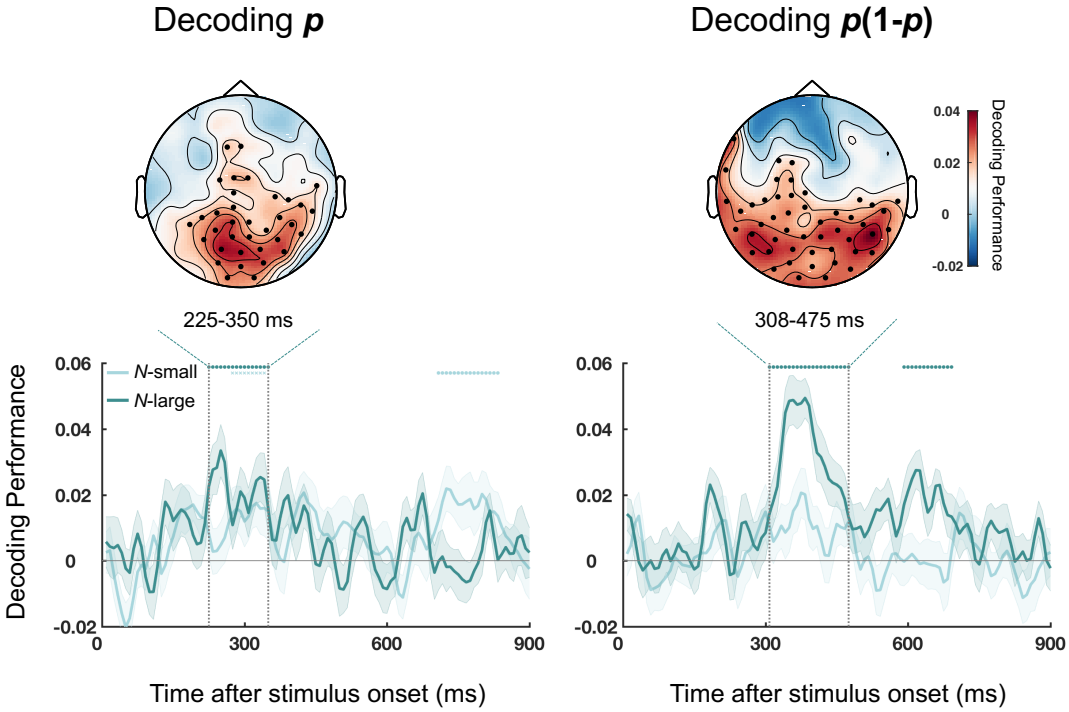

Figure 4. Results of decoding analyses. Main plots: Time-resolved decoding performance over different time lags for *p* (left panel) and *p(1–p)* (right panel), separately for the *N*-small (light green) and *N*-large (dark green) conditions. Shadings denote SEM across subjects. Dots and xs above the curves indicate time lags that had above-chance decoding performance (cluster-based permutation tests; dot: $P_{cluster}<0.01$, x: $0.01\leq P_{cluster}<0.05$). The decoding performance for *p* peaked at around 327 ms after stimulus onset, while that for *p(1–p)* peaked at around 419 ms. **Insets**: Topography of spatial decoding performance for the time window (highlighted by the funnel-shaped dashed lines) that contained the peak decoding performance in the time-resolved decoding analysis. Solid black dots indicate the sensor locations with above-chance decoding performance ($P_{FDR} < 0.05$). The encodings of *p* and *p(1–p)* involved overlapping parietal-occipital regions, including the parietal region where $\Delta\beta$ and $\Delta\hat{\gamma}$ were positively correlated (Figure 3).

Discussion

We used MEG to investigate probability distortions in a relative-frequency estimation task and found that the human brain encodes not only the task-relevant relative frequency but also its task-irrelevant representational uncertainty. The neural encoding of the representational uncertainty occurs as early as approximately 400 ms and influences the probability distortion in the subsequent behavioral judgment. These results suggest that the human brain automatically and quickly quantifies the uncertainty inherent in probability information. As we discuss below, our findings provide neural basis for the theoretical hypothesis that probability distortion is a consequence of uncertainty compensation and, more generally, shed light on the functional role of confidence (estimate of uncertainty) during the process of judgment and decision making.

Neural computations underlying probability distortions

Humans show highly similar probability distortions in tasks involving probability or relative frequency, with the subjective probability typically an inverted-*S*-shaped function of the objective probability (Zhang & Maloney, 2012). Why do people distort probability information in such a systematic way? Though inverted-*S*-shaped distortions of probability had been found in the activity of a few brain regions (Hsu, Krajbich, Zhao, & Camerer, 2009; Tobler, Christopoulos, O'Doherty, Dolan, & Schultz, 2008), the neural computations involved in transforming objective probabilities to subjective probabilities were largely unknown.

Several theories (Fennell & Baddeley, 2012; Martins, 2006; See & Fox, 2006; Zhang et al., 2019) rationalize probability distortion as a consequence of compensating for representational uncertainty that is in accordance with the framework of Bayesian Decision Theory (Maloney & Zhang, 2010). In brief, the brain will discount an internal representation of probability according to the level of uncertainty so that it can achieve a more reliable estimate of the objective probability. The higher the uncertainty associated with a representation, to a greater extent the representation is discounted. By modeling human subjects' probability distortion behaviors in two representative tasks—relative-frequency estimation and decision under risk—Zhang et al. (2019) found evidence that the representational uncertainty of probability or relative frequency p is proportional to $p(1–p)$.

What we found in the present study is neural evidence for the encoding of $p(1–p)$ during the evaluation of p. For the first time, we have shown that representational uncertainty $p(1–p)$ is not just virtual quantities assumed in the computational models of probability distortions but is really computed in the brain.

Representational uncertainty and confidence

The representational uncertainty that concerns us lies in the internal representation of probability or relative-frequency. One concept that may occasionally coincide with such epistemological uncertainty but is entirely different is outcome uncertainty, which is widely studied in the literature of decision under risk (Fiorillo, Tobler, & Schultz, 2003; Tobler, O'Doherty, Dolan, & Schultz, 2007), prediction error (Fiorillo et al., 2003)

and surprise (Preuschoff, 't Hart, & Einhäuser, 2011; van Lieshout, Vandenbroucke, Müller, Cools, & de Lange, 2018). For example, for a gamble with probability p to receive x_1 and probability $1–p$ to receive x_2, the variance of the outcomes is proportional to $p(1–p)$. In the present study, the relative-frequency estimation task we chose to investigate has nothing to do with uncertain outcomes and thus avoids the possible confusion of outcome uncertainty with representational uncertainty.

Representational uncertainty is closely related to confidence, except that it is not necessarily bound to any specific judgment or decision (Pouget, Drugowitsch, & Kepecs, 2016). Specifically, the encoding of representational uncertainty may be understood as a special case of the "second-order valuation" proposed by Lebreton et al. (2015). As described earlier, Lebreton et al. (2015) found a principled relationship between an overt numerical judgment and the individual's confidence about the judgment, with the latter being a quadratic function of the former. They showed in fMRI studies that such quadratic form—as a proxy to confidence—is automatically encoded in vmPFC, even when confidence rating is not explicitly required of. Using the same quadratic form as the proxy, here we found automatic encoding of the representational uncertainty of relative frequency during the tracking of a visual sequence. Our findings go beyond previous work and contribute to the confidence literature in the following three aspects.

First, we found that the encoding of representational uncertainty is even more automatic than the automatic encoding of confidence considered by Lebreton et al. (2015). In our experiment, subjects were asked to track the relative frequency in the

stimulus sequence but made no overt judgment except for the last display. Moreover, displays were refreshed every 150 ms, apparently leaving no time for deliberation.

Second, we resolved the temporal course of the encoding of representational uncertainty, which would be inaccessible under the low temporal resolution of fMRI. In particular, representational uncertainty is encoded as early as 400 ms upon stimulus onset, approximately 100 ms after the encoding of relative frequency itself. The occipital-parietal region is involved in the processing of both relative-frequency and its representational uncertainty but during different time windows. The fast encoding of representational uncertainty we found echoes recent psychophysiological findings in humans (Gherman & Philiastides, 2015, 2018; Zizlsperger, Sauvigny, Händel, & Haarmeier, 2014) and non-human primates (Kiani & Shadlen, 2009) that the confidence encoding for perceptual judgments can be detected in the brain far before confidence rating and even prior to the overt judgment or decision making.

Third, our findings shed light on the functional importance of confidence encoding in information processing (Bahrami et al., 2010; Koriat, 2015; Pouget et al., 2016) that is still not well understood but receives growing attention. In particular, we asked how the neural encoding of representational uncertainty may relate to probability distortion. By comparing experimental conditions under which the same individuals' distortions of relative frequency differed, we found that a relatively stronger response to representational uncertainty in the parietal region corresponds to a shallower slope of distortion. That is, the automatic, fast encoding of representational uncertainty is not just a *post hoc* evaluation but regulates probability distortion.

Methodology implications and future directions

The “steady-state responses” (SSR) technique (Norcia, Appelbaum, Ales, Cottereau, & Rossion, 2015)—using rapid periodic stimulus sequence to entrain brain responses—has been widely used with EEG/MEG to investigate low-level perceptual processes, which sacrifices temporal information to increase the signal-to-noise ratio of detecting the automatic brain responses to the stimuli. In our experiment, we constructed periodic *p* or *p(1–p)* sequences and showed that SSR can also be used to reveal the brain’s automatic responses to these more abstract variables. Moreover, we demonstrated that it is possible to perform time-resolved decoding for the other aperiodic variable embedded in the same sequence. In this way, our experimental design exploited the advantages of both the SSR and time-resolved decoding techniques. Such design can be useful for a broad range of problems where the processing of two or more variables needs to be dissociated.

In conclusion, we found a correlation between the neural encoding of representational uncertainty and the slope of behavioral probability distortion. One question is whether there is a causal relationship between the two. A second future question is how our findings for relative frequency can apply to probability distortions with other types of probability information.

Materials and Methods

Ethics Statement

The MEG experiment had been approved by the Institutional Review Board of School of Psychological and Cognitive Sciences at Peking University (#2015-03-13c). Subjects gave written informed consent in accordance with the Declaration of Helsinki.

Subjects

Twenty-two human subjects (aged 18–27, 13 female) participated. All of them had normal or corrected-to-normal vision and passed the Farnsworth-Munsell 100 Hue Color Vision Test (Farnsworth, 1943).

Experimental Procedures

Apparatus. Subjects were seated approximately 86 cm in front of a projection screen (Panasonic PT-DS12KE: 49.6 × 37.2 cm, 1024 × 768 pixels, 60-Hz refresh rate) inside the magnetically shielded room. Stimuli were controlled by a Dell computer using Matlab and PsychToolbox-3 (Brainard, 1997; Pelli, 1997). Subjects' behavioral responses were recorded by an MEG-compatible mouse system (FOM-2B-10B from Nata technologies) and their brain activities by a 306-channel MEG system (see MEG Acquisition and Preprocessing for details).

Task. Each trial started with a white fixation cross on a blank screen for 600 ms, following which a sequence of displays of yellow and cyan dots was presented on a

gray background at a rate of 150 ms per display (Figure 1A). Subjects were asked to fixate on the central fixation cross and track the relative-frequency of each display. After the sequence ended for 1000 ms, a horizontal scale (0 to 100%) appeared on the screen and subjects were required to click on the scale to indicate the relative-frequency of yellow (or cyan) dots on the last display. Half of the subjects reported relative frequency for yellow dots and half for cyan dots.

To encourage subjects to pay attention to each display, one out of six trials were catch trials whose duration followed a truncated exponential distribution (1–6 s, mean 3 s), such that almost each display could be the last display. The duration of formal trials was 6 or 6.15 s. Only the formal trials were submitted to behavioral and MEG analyses.

On each display, all dots were randomly scattered without overlapping within an invisible circle that subtended a visual angle of 12°. The visual angle of each dot was 0.2° and the center-to-center distance between any two dots was at least 0.12°. The two colors of the dots were isoluminant (CIE_{yellow}=[43, 19.06, 52.33], CIE_{cyan}=[43, -15.49, -23.72]). Isoluminant dark gray pixels (CIE_{gray}=[43, 0, 0]) were filled on the gray background ($CIE_{background}$=[56.5, 0, 0]) between the dots as needed to guarantee each display had equal overall luminance, which prevented luminance from being a confounding factor for any abstract variables of interest.

Design. We adopted a steady-state response (SSR) design, which could achieve a higher signal-to-noise ratio than the conventional event-related design (Norcia et al., 2015). The basic idea was to vary the value of a variable periodically at a specific

temporal frequency and to observe the brain activities at the same frequency as an idiosyncratic response to the variable. The variables of most interest in the present study were relative frequency p and its representational uncertainty quantified by $p(1-p)$.

For half of the trials (referred as the *P*-cycle condition), the value of p in the sequence was chosen from uniform distributions with the ranges alternatively being (0.1, 0.5) and (0.5, 0.9) so that the sequence of p formed cycles of 3.3 Hz. For the other half trials (referred as the *U*-cycle condition), the value of p was chosen alternatively from (0.1 0.3) U (0.7, 0.9) and (0.3, 0.7) so that the sequence of $p(1-p)$ (i.e. proxy for uncertainty) was in cycles of 3.3 Hz. That is, the two cycle conditions had matched individual displays but differed in which variable formed a periodic sequence. Accordingly, the $p(1-p)$ sequence in the *P*-cycle condition and the p sequence in the *U*-cycle condition were aperiodic, which had no autocorrelations (Figure S2).

The total number of dots on a display (numerosity, denoted N) was varied across trials as well as across individual displays of the same trial. The value of N for a specific display was independent of the value of p and generated as a linear transformation of a beta random variable following Beta(0.1, 10) so that the number of dots in each color (pN or $(1-p)N$), a possible confounding variable of p, would only have moderate correlations (Pearson's $|r| < 0.5$) with p. The linear transformation was chosen to locate N to the range of [10, 90] for half of the trials (the *N*-small condition) and to [100, 900] for the other half (the *N*-large condition).

The numbers of dots in each color was *pN* or (1–*p*)*N* rounded to the nearest integer. As the result of rounding, the actual relative-frequency was slightly different from the originally chosen *p*. In data analysis, we would use *p* to denote the actual relative-frequency that was presented to subjects.

To summarize, there were 2 (*P*-cycle vs. *U*-cycle) by 2 (*N*-small vs. *N*-large) experimental conditions, with all conditions interleaved and each condition repeated for 6 trials in each block. Each subject completed 10 blocks of 24 trials, resulting in 50 formal trials and 10 catch trials for each condition. Before the formal experiment, there were 32 practice trials (the first 20 trials consisted of one display and the following 12 trials had the same settings as the formal experiment) for subjects to be familiar with the procedure. No feedback was available during the experiment.

Behavioral Analysis

Measures of probability distortion. According to Zhang and Maloney (2012), inverted-*S*- or *S*-shaped probability distortions can be well captured by the LLO model

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

where p and $\pi\left(p\right)$ respectively denote the objective and subjective probability or relative-frequency, $\lambda\left[p\right]=\log\frac{p}{1-p}$ is the log-odds transformation, and ε_λ is Gaussian error on the log-odds scale with mean 0 and variance σ_λ^2. The γ, p_0, and σ_λ are free parameters. The parameter γ is the slope of distortion, with $\gamma<1$ corresponding to inverted-*S*-shaped distortion, $\gamma=1$ to no distortion, and $\gamma>1$ to *S*-shaped distortion. The parameter p_0 is the crossover point where $\pi\left(p\right)=p$.

For each subject and condition, we fit the reported relative-frequency to the LLO model (Eq. 2) and used the estimated $\hat{\gamma}$ and $\hat{p}_0$ as the measures of relative-frequency distortions.

Response time. The response time (RT) was defined as the interval between the response screen onset and subjects' first mouse move. The RT of the first trial of three subjects was mis-recorded due to technical issues and was excluded from further analysis. We divided all trials evenly into 5 bins based on the to-be-reported *p*, and computed the mean *p* and mean RT for each bin. A 5 (*p* bins) by 2 (numerosity conditions) repeated-measures ANOVA was performed on the mean RTs. A similar binning procedure was applied to *p*(1-*p*) to visualize the relationship of RT to *p*(1-*p*).

MEG Acquisition and Preprocessing

Subjects' brain activity was recorded by a 306-channel whole-head MEG system (Elekta-Neuromag, 102 magnetometers and 102 pairs of orthogonal planar gradiometers). Head position was measured before each block by an isotrack polhemus system with four head position indicator coils (two on the left and right mastoid, the other two on the left and right forehead below the hairline). Subjects whose between-block head movement exceeded 3 mm would be excluded from further analysis. Horizontal and vertical Electro-oculograms were recorded to monitor eye-movement artifacts. Sampling rate was set to be 1000 Hz and an analog band-pass filter from 0.1 to 330 Hz was applied. Maxwell filtering was used to minimize external magnetic interference and to compensate for head movements

(Samu Taulu & Kajola, 2005; S. Taulu & Simola, 2006).

Standard preprocessing procedures were applied using Matlab 2016b and the FieldTrip package (Oostenveld, Fries, Maris, & Schoffelen, 2011). The MEG data of each block was first low-pass filtered below 20-Hz and then segmented into epochs of 7.6 s relative to trial onset (–0.6 s to 7 s). Independent component analysis (ICA) was applied to aggregated epoch data to remove artifacts including blinks, eye movements, breaths and heart activity. No subject was excluded for excessive head movement or other artifacts. For two subjects, the first trial of one block was excluded due to the failure of synchronization between stimulus onset and MEG recording onset at the beginning of the block.

Phase Coherence Analysis

Given a periodically-changed stimulus sequence (*p* in *P*-cycle or *p*(1–*p*) in *U*-cycle), stimulus-evoked brain responses imply that the phase difference between the stimulus and response time series should be coherent across trials at the stimulus frequency (Summerfield & Mangels, 2004). The phase coherence across trials is defined as (Kramer & Eden, 2016):

$$\kappa = \frac{\left|\overline{P}_{rs}\right|}{\sqrt{\overline{P}_{ss}}\sqrt{\overline{P}_{rr}}} \quad , \tag{3}$$

where $\overline{P}_{ss}$ and $\overline{P}_{rr}$ respectively denote the trial-averaged power spectrum for stimulus and response time series, and $\left|\overline{P}_{rs}\right|$ denotes the magnitude of the trial-averaged cross-spectrum between stimuli and responses. The value of κ for any specific frequency is between 0 and 1, with larger κ indicating stronger

phase-locked responses.

For a specific variable (*p* or *p*(1–*p*)), we computed the phase coherence between stimulus sequence and MEG signals separately for each subject, each magnetometer and each cycle and numerosity condition. In particular, we down-sampled epoched MEG time series to 300 Hz, applied zero padding and Hanning window to single trials, and used Fast Fourier Transformation (FFT) to calculate power spectrum. The resulting frequency resolution was 0.03 Hz.

To evaluate the chance-level phase coherence, we shuffled the MEG time series across different time points within each trial and computed phase coherence for the permutated time series. This permutation procedure was repeated for 500 times to produce a distribution of chance-level phase coherences.

Decoding Analysis

Time-resolved decoding. For an aperiodic stimulus sequence (*p*(1–*p*) in *P*-cycle or *p* in *U*-cycle), we could infer the brain's encoding of the stimulus variable at a specific time lag to stimulus onset by examining how well the value of the variable could be reconstructed from the neural responses at the time lag. In particular, we performed a time-resolved decoding analysis using regression methods (Crosse, Liberto, Bednar, & Lalor, 2016; Gonçalves, Whelan, Foxe, & Lalor, 2014; Lalor, Pearlmutter, Reilly, McDarby, & Foxe, 2006; Wyart, Gardelle, Scholl, & Summerfield, 2012):

$$S(t) = \sum_{n} g(\tau, n) R(t+\tau, n) + \varepsilon(t) \quad , \tag{4}$$

where $S(t)$ denotes the value of the stimulus that starts at time *t*, $g(\tau, n)$ denotes

the to-be-estimated decoding weight for channel n at time lag τ, $R(t+\tau,n)$ denotes the response of channel n at time $t+\tau$, and $\varepsilon(t)$ is a Gaussian noise term.

We used the MEG signals of all 306 sensors (102 magnetometers and 204 gradiometers) at single time lags to decode *p* from *U*-cycle trials or *p*(1–*p*) from *P*-cycle trials, separately for each subject, each cycle and numerosity condition, and each time lag between 0 and 900 ms. Epoched MEG time series were first down-sampled to 120 Hz. To eliminate multicollinearity and reduce overfitting, we submitted the time series of the 306 sensors to a principal component analysis (PCA) and used the first 30 components (explaining approximately 95% variance) as the regressors. The decoding weights (i.e. regression coefficients) were estimated for normalized stimuli and responses using the L2-rectified regression method implemented in the mTRF toolbox (Crosse et al., 2016) , with the ridge parameter set to 1.

We used a leave-one-out cross-validation procedure to evaluate the predictive power of decoding performance as the following. Each time one of the 50 trials in question (or 49 trials in the case of trial exclusion) served as the test set and the remaining trials as the training set. The decoding weights estimated from the training set were applied to the test set, for which Pearson's *r* was calculated between the predicted and the ground-truth stimulus sequence. Such computation was repeated for each trial as the test set and the averaged Pearson's *r* was used as the measure for decoding performance.

We identified time windows that had above-chance decoding performance using cluster-based permutation tests (Maris & Oostenveld, 2007) as follows. Adjacent time lags with significantly positive decoding performance at the uncorrected significance level of .05 by right-sided one-sample *t* tests were grouped into clusters and the summed *t*-value across the time lags in a cluster was defined as the cluster-level statistic. We randomly shuffled the stimulus sequence of each trial, performed the time-resolved decoding analysis on the shuffled sequence and recorded the maximum cluster-level statistic. This procedure was repeated for 500 times to produce a reference distribution of chance-level maximum cluster-level statistic, based on which we calculated the *P* value for each cluster in real data. This test effectively controls the Type Ⅰ error rate in situations involving multiple comparisons. For each subject, we defined the median of all time points for which the decoding performance exceeded the 95% percentile of the distribution as the time lag of the peak (Marti, King, & Dehaene, 2015).

Spatial decoding. For the conditions and time windows that had above-chance performances in the time-resolved decoding analysis, we further performed a spatial decoding analysis at individual sensor location to locate the brain regions that are involved in encoding the variable in question. For a specific senor location, the MEG signals at its three sensors (one magnetometer and two gradiometers) within the time window were used for decoding. The time window that enclosed the peak in the time-resolved decoding was 125-ms and 167-ms wide respectively for *p* and *p*(1-*p*)

under *N*-large condition. The decoding procedure was similar to that of the time-resolved decoding described above, except that the number of PCA component used was 19 and 24 respectively for decoding *p* and *p*(1-*p*), which explained approximately 99% variance of the original 48 (16 time lags × 3 sensors) and 63 (21 time lags × 3 sensors) channels.

Statistical Analysis

For phase coherence or decoding performance, grand averages across subjects were computed and permutation tests described above were used to assess their statistical significance over chance (Maris & Oostenveld, 2007). The false discovery rate (*FDR*) method (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001) was used for multiple comparison corrections whenever applicable. For the phase coherence spectrum averaged across magnetometers, *FDR* corrections were performed among the 235 frequency bins within 0–7 Hz. For the phase coherence topography at 3.33 Hz, *FDR* corrections were performed among all 102 magnetometers.

References

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology, 46*(2), 81-86.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329*, 1081-1085.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*(4), 1165-1188.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436.

Crosse, M. J., Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *frontiers in Human Neuroscience, 10*, 604-617.

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.

Farnsworth, D. (1943). The Farnsworth-Munsell 100 Hue and Dichotomous Test for Color Vision. *Journal of the Optical Society if America, 33*(10), 568-578.

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

878-887. doi:10.1037/a0029346

Ferrari-Toniolo, S., Bujold, P. M., & Schultz, W. (2019). Probability distortion depends on choice sequence in Rhesus Monkeys. *The Journal of Neuroscience, 39*(15), 2915-2929.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299*(5614), 1898-1902.

Gherman, S., & Philiastides, M. G. (2015). Neural representations of confidence emerge from the process of decision formation during perceptual choices. *NeuroImage, 106*, 134-143.

Gherman, S., & Philiastides, M. G. (2018). Human VMPFC encodes early signatures of confidence in perceptual decisions. *eLIFE, 7*, e38293. doi:https://doi.org/10.7554/eLife.38293

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review, 98*(4), 506-528.

Gonçalves, N. R., Whelan, R., Foxe, J. J., & Lalor, E. C. (2014). Towards obtaining spatiotemporally precise responses to continuous sensory stimuli in humans: A general linear modeling approach to EEG. *NeuroImage, 97*, 196-205.

Gonzalez, R., & Wu, G. (1999). On the Shape of the Probability Weighting Function. *Cognitive Psychology, 38*, 129-166.

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

Huang, Q., Jia, J., Han, Q., & Luo, H. (2018). Fast-backward replay of sequentially memorized items in humans. *eLIFE, 7*, e35164. doi:https://doi.org/10.7554/eLife.35164

Jia, J., Liu, L., Fang, F., & Luo, H. (2017). Sequential sampling of visual objects during sustained attention. *PloS Biology, 15*(6), 1-19.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*(8), 759-763.

Koriat, A. (2015). Metacognition: decision-making processes in self-monitoring and self-regulation. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (Vol. 1, pp. 356-379). Malden, MA: Wiley Blackwell.

Kramer, M. A., & Eden, U. T. (2016). *Case studies in neural data analysis: A guide for the practicing neuroscientist* (Vol. 11). the United States of America: Massachusetts Institute of Technology.

Lalor, E. C., Pearlmutter, B. A., Reilly, R. B., McDarby, G., & Foxe, J. J. (2006). The VESPA: A method for the rapid estimation of a visual evoked potential. *NeuroImage, 32*, 1549-1561.

Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valutation signal. *Nature Neuroscience, 18*(8), 1159-1167.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judge Frequency of Lethal Events. *Journal of Experimental Psychology: Human*

Learning and Memory, 4(6), 551-578.

Liu, L., Wang, F., Zhou, K., Ding, N., & Luo, H. (2017). Perceptual integration rapidly activates dorsal visual pathway to guide local processing in early visual areas. *PloS Biology, 15*(11), e2003646.

Luce, R. D. (2000). *Utility of gains and losses: measurement -- theoretical, and experimental approaches* Mahwah, New Jersey, London: Lawrence Erlbaum Associates, Inc.

Maloney, L. T., & Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Research, 50*(23), 2362-2374.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*, 177-190.

Marti, S., King, J.-R., & Dehaene, S. (2015). Time-resolved decoding of two processing chains during dual-task interference. *Neuron, 88*, 1297-1307.

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

Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., & Rossion, B. (2015). The steady-state visual evoked potential in vision research: A review. *Journal of Vision, 15*(6), 1-46.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011*, 1-9.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*(4), 437-442.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience, 19*(3), 366-374.

Preuschoff, K., 't Hart, B. M., & Einhäuser, W. (2011). Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. *frontiers in Neuroscience, 5*(115), 1-12.

See, K. E., & Fox, C. R. (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.

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.

Summerfield, C., & Mangels, J. A. (2004). Coherent theta-band EEG activity predicts item-context binding during encoding. *NeuroImage, 24*, 692-703.

Taulu, S., & Kajola, M. (2005). Presentation of electromagnetic multichannel data: The signal space separation method. *Journal of Applied Physics, 97*, 124905 124901-124910.

Taulu, S., & Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine and Biology, 51*, 1759-1768.

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., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *J Neurophysiol, 97*, 1621-1632.

Tversky, A., & Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty, 5*, 297-323.

van Lieshout, L. L. F., Vandenbroucke, A. R. E., Müller, N. C. J., Cools, R., & de Lange, F. P. (2018). Induction and relief of curiosity elicit parietal and frontal activity. *The Journal of Neuroscience, 38*(10), 2579-2588.

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.

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

Wyart, V., Gardelle, V. d., Scholl, J., & Summerfield, C. (2012). Rhythmic Fluctuations in Evidence Accumulation during Decision Making in the Human Brain. *Neuron, 76*, 847-858.

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

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.

Zhang, H., Ren, X., & Maloney, L. T. (2019). The bounded rationality of probability distortion: the bounded log-odds model. *bioRxiv*. doi:https://doi.org/10.1101/662429

Zizlsperger, L., Sauvigny, T., Händel, B., & Haarmeier, T. (2014). Cortical representations of confidence in a visual perceptual decision. *Nature Communications, 5*, 3940.

Supplement S1. Fitting the Bounded Log-Odds (BLO) model to subjects' $\pi(p)$

Please refer to Zhang, Ren, & Maloney (2019) for a full description of the BLO model, which we briefly formulize below. BLO assumes that any probability or relative frequency, *p*, is internally represented as a linear transformation of log-odds,

$$\lambda(p) = \log \frac{p}{1-p} \ . \tag{S1}$$

A bounding operation

$$\Gamma[\lambda] = \begin{cases} \Delta^-, & \lambda < \Delta^- \\ \lambda, & \Delta^- \le \lambda \le \Delta^+ \\ \Delta^+, & \lambda > \Delta^+ \end{cases} \tag{S2}$$

is defined to confine the representation of log-odds λ to the interval $[\Delta^-, \Delta^+]$, where Δ^- and Δ^+ are free parameters. $\Lambda(p) = \Gamma(\lambda[p])$ is referred as "truncated log-odds".

In our relative-frequency judgment task, we assume representational uncertainty partly arises from random variation of sampling. Suppose subjects may not access to all of the dots in a display. For a display of N dots, a sample of n_s dots is randomly drawn from the display without replacement and used to infer the relative frequency. The resulting variance of $\hat{p}$ is thus (Cochran, 1977; Zhang et al., 2019):

$$V(\hat{p}) \ = \frac{p(1-p)}{n_s} \frac{N-n_s}{N-1} \ . \tag{S3}$$

Here n_s is modeled as an increasing function of numerosity:

$$n_s = b + N^a \ , \tag{S4}$$

where $a \ge 0$ and $b \ge 0$ are free parameters.

To compensate for this representational uncertainty, the internal representation $\Lambda(p)$ further undergoes a non-linear contraction or expansion around a fixed point Λ_0 on the log-odds scale so that the log-odds of the reported relative frequency

$$\lambda\left[\pi(p)\right] = \eta\left[\ \omega_p \Lambda(p) + \left(1-\omega_p\right)\Lambda_0\ \right] + \varepsilon_\lambda\ , \tag{S5}$$

where $\omega_p = \dfrac{1}{1+\alpha V\left(\hat{p}\right)}$ is a measure of the reliability of the internal representation, and ε_λ is Gaussian error on the log-odds scale with mean 0 and variance σ_λ^2. The α, η, Λ_0, and σ_λ are free parameters. Along with the Δ^-, Δ^+, a and b, the BLO model has 8 parameters. Finally, the scaling factor η adjusts the bounded interval to the fixed Thurstone scale (Thurstone, 1927; Zhang et al., 2019).

For each subject, we fit the BLO model to the subject's $\pi(p)$ in each trial using maximum likelihood estimates. The *fminsearchbnd* (J. D'Errico), a function based on *fminsearch* in MATLAB (MathWorks), was used to search for the parameters that minimized negative log likelihood.

To summarize how the slope of distortion changed with numerosity, for each subject, we divided the trials of each numerosity condition equally into two sub-conditions according to the value of *N* and then estimated $\hat{\gamma}$ for each of the four sub-conditions (Figure S1). The pattern of $\hat{\gamma}$ in real data was well predicted by the fitted BLO model (based on 500 times of simulations).

We also compared the goodness-of-fit of the BLO model with that of the LLO model (Eq. 2 in the main text), which separately fit the slopes of the *N*-small and *N*-large conditions (4 parameters in total, $\gamma_{N\text{-small}}$, $\gamma_{N\text{-large}}$, p_0 and ε). According to the Akaike information criterion with a correction for sample sizes (AICc) (Akaike, 1974; Hurvich & Tsai, 1989), the BLO model outperformed the LLO model (the summed AICc difference across subjects between the LLO and BLO models was 285.33).

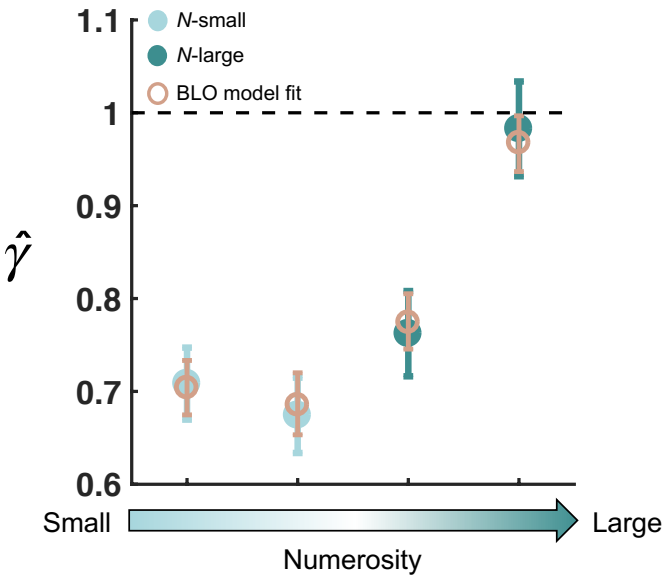

Figure S1. Estimated slopes of distortion: data versus BLO model fits. According to the numerosity of the last display, the trials in the *N*-small and *N*-large conditions were further divided into two equal halves, whose slopes of distortions, $\hat{\gamma}$, were estimated separately. Light and dark green dots denote data. Orange circles denote BLO fits. Error bars for data and BLO fits respectively denote SEM and the 95% confidence interval. The BLO model, which assumes compensation for representational uncertainty in the $p(1\text{-}p)$ form, well captures how $\hat{\gamma}$ varies with numerosity.

Supplement S2. Autocorrelation analysis for periodic and aperiodic stimulus sequences

We applied an autocorrelation analysis to the stimulus sequences in each trial to confirm their periodicity or aperiodicity. Autocorrelations were calculated for individual trials and then averaged across trials in each condition. As designed, the *p* sequence in *P*-cycle trials and the *p(1–p)* sequence in *U*-cycle trials had periodic autocorrelations, while the *p* sequence in *U*-cycle trials and the *p*(1-*p*) sequence in *P*-cycle trials were free of autocorrelations (Figure S2).

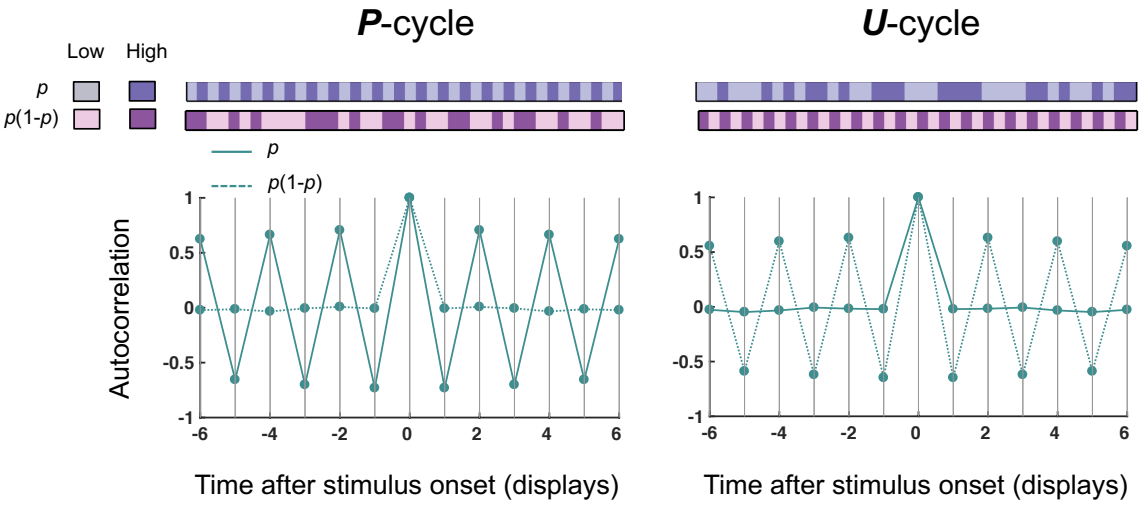

Figure S2. **Autocorrelations of stimulus sequences as functions of time lags.** Left: the *P*-cycle condition. Right: the *U*-cycle condition. Solid and dashed lines are respectively for the *p* and *p*(1-*p*) sequences. The stimulus sequences used for autocorrelation calculations were from the example subject whose behavior results are showed in Figure 1C. All other subjects had similar autocorrelation patterns.

Supplement S3. Phase coherence and time-resolved decoding analyses for confounding factors

We considered the following eight variables in the stimuli as potential confounding factors, which may covary with *p* or *p(1–p)* across displays.

(1) *N*: the total number of dots (numerosity) in the display, which by design was independent of both *p* and *p*(1–*p*).

(2) N_t: the number of dots in the target color in the display, that is, $N_t = Np$.

(3) N_o: the number of dots in the other color in the display, that is, $N_o = N(1-p)$.

(4) *AvgLumi*: the mean luminance of the display, which was kept constant across different displays and different trials and was thus independent of both *p* and *p*(1–*p*).

(5) *vCIE-L**: the color variance of the display on the L* dimension in the CIELAB color space, where L* ranges from black to white.

(6) *vCIE-a**: the color variance of the display on the a* dimension in the CIELAB color space, where a* ranges from green to red.

(7) *vCIE-b**: the color variance of the display on the b* dimension in the CIELAB color space, where b* ranges from blue to yellow.

(8) *M-contrast*: the *Michelson* luminance contrast (Wiebel, Singh, & Maertens, 2016).

According to our design, most of these variables had negligible correlations with *p* and *p(1–p)* and none of the absolute values of correlation coefficients exceeded 0.53 (Pearson's *r*, Figure S3A). The numerosity (*N*) had considerable correlations (Pearson's $|r| > 0.79$) with all other confounding factors except for the mean luminance (*AvgLumi*), which was constant across displays. We performed similar phase coherence (Figure S3B) and time-resolved decoding analyses (Figure S3C) for these variables as we did for *p* and *p(1–p)* but obtained very different patterns, which suggests that the findings we reported for *p* and *p(1–p)* in the main text are unlikely to be effects of confounding factors. The patterns of all these confounding factors expect for *AvgLumi* were almost identical, which may have the same

origins such as the automatic encoding of numerosity (Fornaciai, Brannon, Woldorff, & Park, 2017; Park, DeWind, Woldorff, & Brannon, 2016).

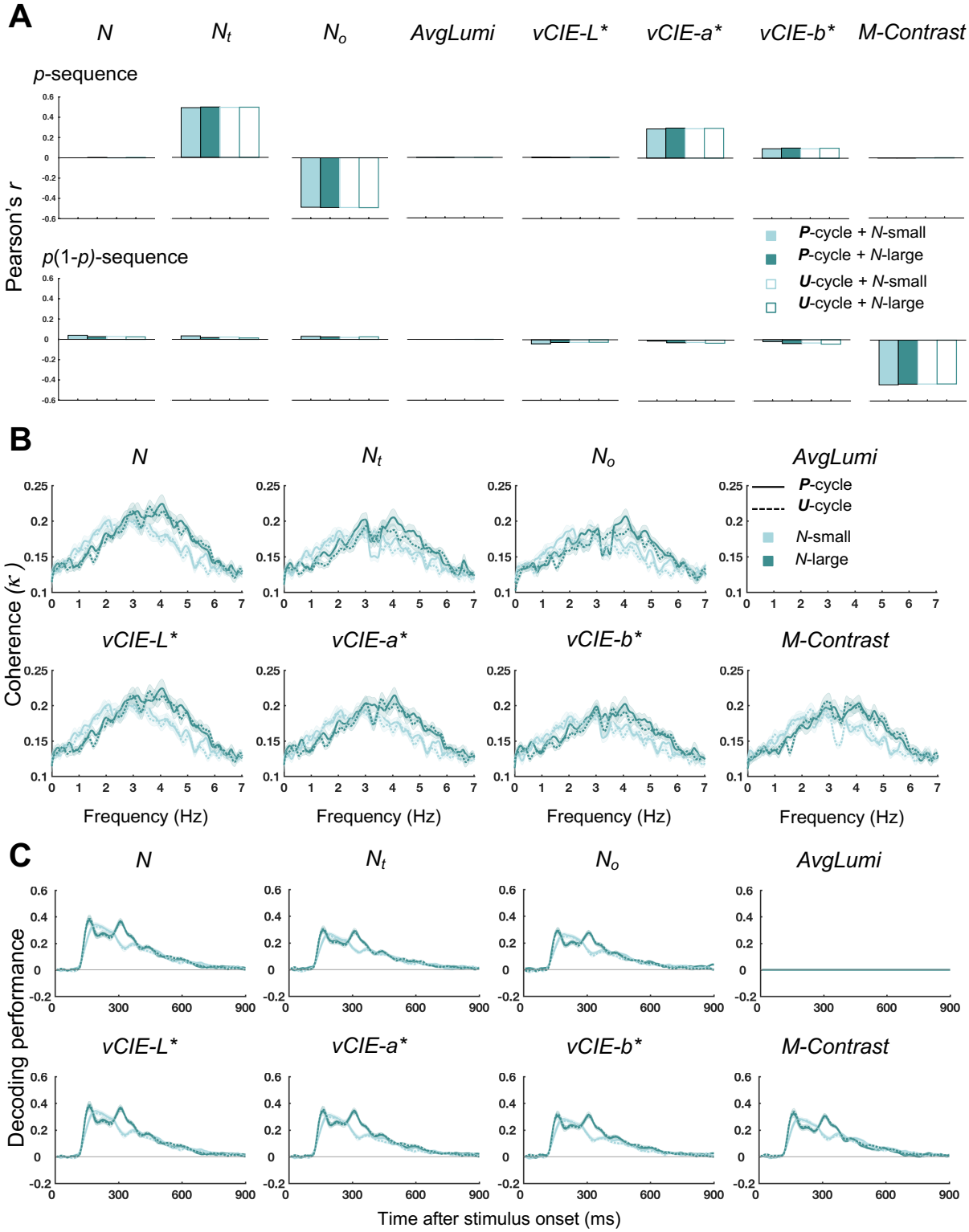

Figure S3. **Analysis for the confounding variables. (A) Correlation analysis.** Top row: Pearson's correlations between the *p* sequence and the eight confounding variables (*N*, N_t, N_o, *AvgLumi*, *vCIE-L**, *vCIE-a**, *vCIE-b** and *M-contrast*) separately for the four cycle and numerosity conditions. Bottom row: Pearson's correlations between the *p*(1–*p*) sequence and the eight confounding variables for each condition. The results confirmed our design that all of the confounding variables had negligible or moderate correlations with *p* and *p(1–p)*. **(B) Phase coherence analysis.** Grand-averaged phase coherence spectrum between the sequences of the eight confounding variables and neural responses from magnetometers. The observed phase coherence spectra of the confounding variables had different patterns than those of *p* or *p*(1–*p*). See the Supplement S4 for the possible cause of the "bumps" between 0 and 6.67 Hz. **(C) Time-resolved decoding analysis.** The same time-resolved decoding procedures as Figure 4 were applied to the eight confounding variables. For **(A)**–**(C)**, light and dark green respectively code the *N*-small and *N*-large conditions. For **(B)** and **(C)**, solid and dashed lines respectively denote the *P*-cycle and *U*-cycle conditions. Shadings denote SEM across subjects.

Supplement S4. Explaining the "bumps" in the phase coherence spectra of the confounding factors

As shown in the Figure S3B, there was a "bump" spanning across 0–6.67 Hz in the phase coherence spectrum of *N*. So did all of the confounding variables we considered above that had high correlations with *N*. Though also centering at 3.33 Hz, these "bumps" were much wider than the 3.33 Hz peak we observed for the *p* sequence in the *P*-cycle condition or the *p(1–p)* sequence in the *U*-cycle condition. Besides, the "bumps" were similar across all four cycle and numerosity conditions, which were unlikely due to the periodicity of *p* or *p(1–p)*. We found in a simulation analysis that these "bumps" could simply result from the particular temporal course of the neural responses to the variables.

In particular, we first estimated the temporal course of stimulus-evoked brain responses to *N* using a computational technique known as Temporal Response Function (TRF) analysis (Crosse, Liberto, Bednar, & Lalor, 2016; Gonçalves, Whelan, Foxe, & Lalor, 2014; Lalor, Pearlmutter, Reilly, McDarby, & Foxe, 2006). The TRF for a specific stimulus dimension describes how the brain responds to a unit change in stimulus magnitude at different time lags. The brain is modeled as a linear system, whose response at a specific time point is the summation of the responses to all previous stimuli, that is, a convolution of the stimulus sequence with TRF plus Gaussian random noise:

$$R(t,n) = \sum_{\tau} w(\tau,n) S(t-\tau) + \varepsilon(t,n), \tag{S6}$$

where *t* denotes time, *n* denotes sensor number, τ denotes the time lag to stimulus onset, $R(\bullet)$, $w(\bullet)$, $S(\bullet)$ and $\varepsilon(\bullet)$ respectively denote the magnitude of sensor response, the to-be-estimated TRF, the stimulus sequence, and Gaussian noise term. The stimulus sequence was modeled as the summation of pulses at display onsets. For the TRF analysis, epoched MEG data were down-sampled to 120 Hz. The TRF $w(\tau,n)$ was estimated for a maximum delay of 900 ms using a rectified regression method, with the ridge parameter set to 1. Both the stimulus sequence and sensor responses were normalized before the analysis.

We estimated the TRF for *N* and then convolved it with the *N* sequence and added a Gaussian noise to generate virtual neural responses. We then applied the phase coherence analysis to the virtual neural responses. This simulation was repeated for 500 times, based on which we computed the mean phase coherence spectrum for the simulated data. We found that the resulting phase coherence spectrum resembled that of real data, reproducing the "bump" between 0 and 6.67 Hz (Figure S4).

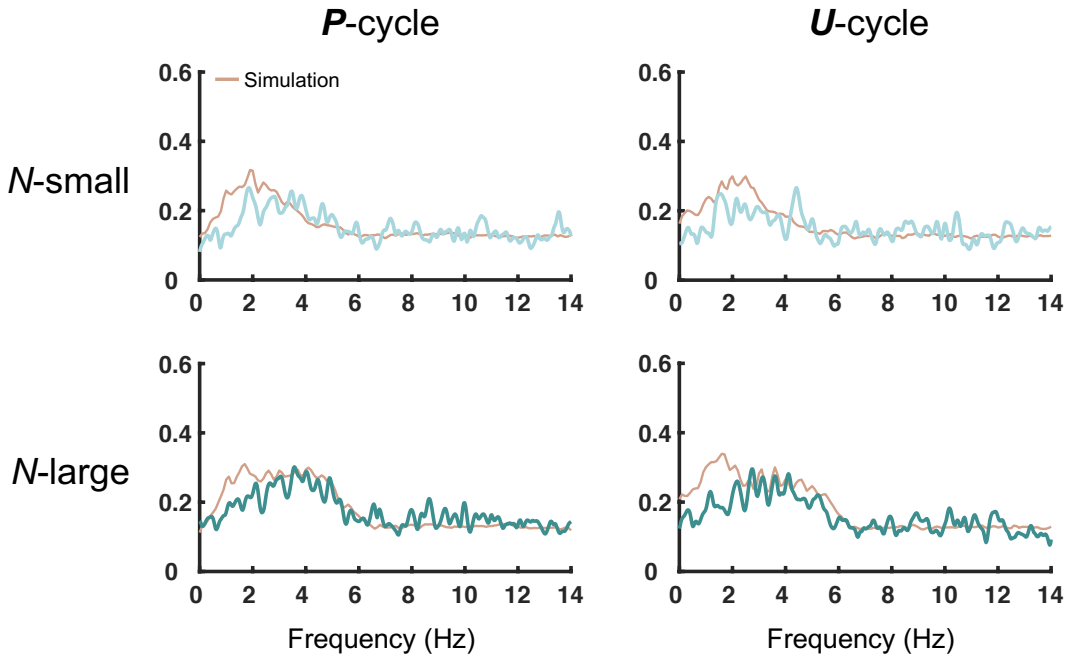

Figure S4. **Explaining the "bumps" in the phase coherence of numerosity.** Simulated neural signals were generated through convolution of TRF estimated from the MEG2521 sensor (located right occipital) with the *N* sequence, perturbed by Gaussian nose whose standard deviation was set to 4. The simulated neural signals were then submitted to the phase coherence analysis (see Materials and Methods). Light and dark green curves denote the phase coherence across magnetometers of real data from one example subject. Orange curves denote the average across 500 simulations. The "bump" across 0–6.67 Hz observed in real data was reproduced in the simulated data.

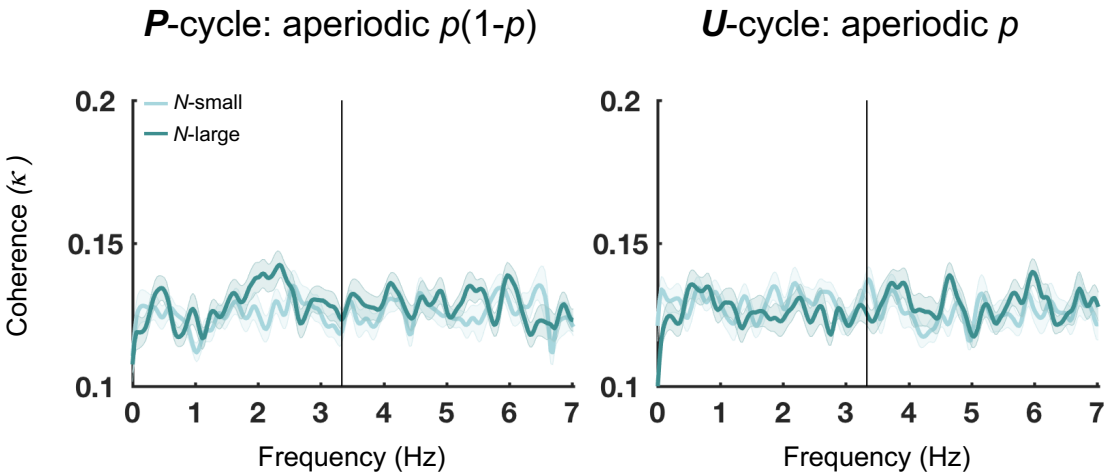

Figure S5. **Phase coherence analysis for the aperiodic sequences.** Grand-averaged phase coherence spectrum across 102 magnetometers. Left: The phase coherence (κ) between the (aperiodic) *p*(1–*p*) sequence and MEG signals in *P*-cycle trials. Right: The phase coherence between the (aperiodic) *p* sequence and MEG signals in *U*-cycle trials. Shadings denote SEM across subjects. The vertical line marks the 3.33 Hz. As expected, none of the aperiodic sequences had phase coherence peaks at 3.33 Hz.

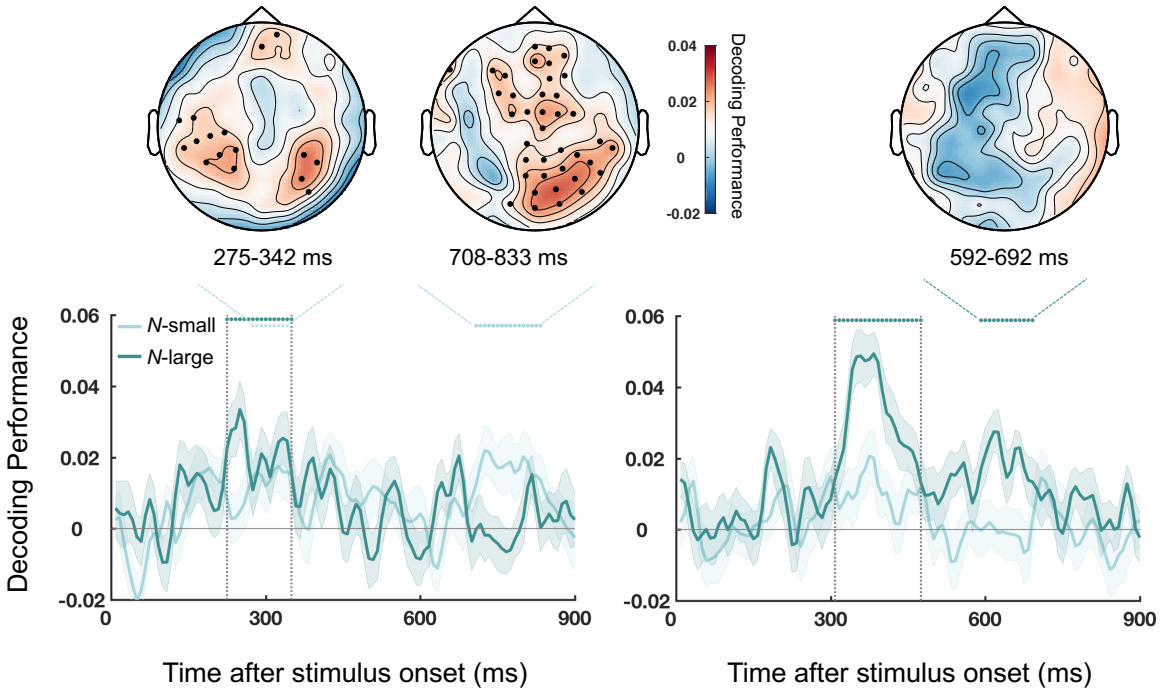

Figure S6. **Topography of spatial decoding performance for the other three statistically significant time windows in the time-resolved decoding analysis.** Black dots mark the sensor locations with above-chance decoding performance (P_{FDR} < 0.05).

Supplement References

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

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

Crosse, M. J., Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *frontiers in Human Neuroscience, 10*, 604-617.

Fornaciai, M., Brannon, E. M., Woldorff, M. G., & Park, J. (2017). Numerosity processing in early visual cortex. *NeuroImage, 157*, 429-438.

Gonçalves, N. R., Whelan, R., Foxe, J. J., & Lalor, E. C. (2014). Towards obtaining spatiotemporally precise responses to continuous sensory stimuli in humans: A general linear modeling approach to EEG. *NeuroImage, 97*, 196-205.

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

Lalor, E. C., Pearlmutter, B. A., Reilly, R. B., McDarby, G., & Foxe, J. J. (2006). The VESPA: A method for the rapid estimation of a visual evoked potential. *NeuroImage, 32*, 1549-1561.

Park, J., DeWind, N. K., Woldorff, M. G., & Brannon, E. M. (2016). Rapid and direct encoding of numerosity in the visual stream. *Cerebral Cortex, 26*, 748-763.

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

Wiebel, C. B., Singh, M., & Maertens, M. (2016). Testing the role of Michelson contrast for the perception of surface lightness. *Journal of Vision, 16*(11), 17, 11-19.

Zhang, H., Ren, X., & Maloney, L. T. (2019). The bounded rationality of probability distortion: the bounded log-odds model. *bioRxiv*. doi:https://doi.org/10.1101/662429