Comparing probabilistic accounts of probability judgments

Derek Powell[1]

[1] Arizona State University, School of Social and Behavioral Sciences

Author Note

This manuscript has not yet been peer-reviewed. This study reports secondary analyses of data a and was not preregistered.

Correspondence concerning this article should be addressed to Derek Powell, 4701 W Thunderbird Rd, Phoenix AZ 85306. E-mail: dmpowell@asu.edu

Abstract

Bayesian theories of cognitive science hold that cognition is fundamentally probabilistic, but people's explicit probability judgments often violate the laws of probability. Two recent proposals, the "Probability Theory plus Noise" (Costello & Watts, 2014) and "Bayesian Sampler" (Zhu et al., 2020) theories of probability judgments, both seek to account for these biases while maintaining that mental credences are fundamentally probabilistic. These theories fit quite differently into the larger project of Bayesian cognitive science, but their many similarities complicate comparisons of their predictive accuracy. In particular, comparing the models demands a careful accounting of model complexity. Here, I cast these theories into a Bayesian data analysis framework that supports principled model comparison using information criteria. Comparing the fits of both models on data collected by Zhu and colleagues (2020) I find the data are best explained by a modified version of the Bayesian Sampler model under which people may hold informative priors about probabilities.

Keywords: probability judgments, Bayesian cognitive science, heuristics and biases

Word count: 5088

Comparing probabilistic accounts of probability judgments

Bayesian theories of cognition offer a unified formal framework for cognitive science (Tenenbaum et al., 2011) that has had remarkable explanatory successes across domains, including in perception (e.g. Kersten et al., 2004), memory (e.g. Anderson, 1991), language (e.g. Xu & Tenenbaum, 2007), and reasoning (e.g. Lu et al., 2012). At the heart of the Bayesian project is the idea that cognition is fundamentally probabilistic: that people reason according to subjective degrees of belief which follow the laws of probability and, in particular, that they are revised in light of evidence according to Bayes' Rule. It is somewhat embarrassing then, that these theories have often been accused of failing to describe human "beliefs" of the simple and everyday sort, such as beliefs like "it will rain tomorrow", "vaccines are safe," or "this politician is trustworthy" (Chater et al., 2020).

Trouble starts as soon as we attempt to measure beliefs. According to Bayesian theories of cognition and epistemology (Jaynes, 2003), the degree to which people believe in various propositions, or their credences, should reflect subjective mental probabilities. So asking people to express beliefs in terms of probability seems only natural.

Unfortunately, people's explicit probability judgments routinely violate the most basic axioms of probability theory. For example, human probability judgments often exhibit the "conjunction fallacy": people will often judge the conjunction of two events (e.g. "Tom Brady likes football and miniature horses") as being more probable than one of the events in isolation (e.g. "Tom Brady likes miniature horses"), a plain and flagrant violation of probability theory (Tversky & Kahneman, 1983). Other demonstrations of the incoherence of probability judgments include disjunction fallacies, subadditivity or "unpacking" effects (Tversky & Koehler, 1994), and a number of others (for an accessible review, see Kahneman, 2013). Altogether these findings have led many researchers to abandon the notion that degrees of belief are represented as probabilities.

Recently however, two groups of researchers have proposed theories of human

probability judgments that account for biases in these judgments while maintaining that mental credences are fundamentally probabilistic (Costello & Watts, 2014; Zhu et al., 2020). Both of these theories build on the increasingly popular notion that a variety of human reasoning tasks are accomplished by a limited process of mental "sampling" from a probabilistic mental model (see also Chater et al., 2020; Dasgupta et al., 2017).[1]

Two probabilistic theories of probability judgment

Costello and Watts (2014, 2016, 2018) have proposed a theory of probability judgment they call the "Probability Theory plus Noise" theory (PT+N). In the PT+N model, mental "samples" are drawn from a probabilistic mental model of events and are then "read" with noise, so that some positive examples will be read as negative and some negative examples read as positive with some probability d. The results are probability judgments reflecting probabilistic credences perturbed by noise. In their model, the probability a mental sample for an event A is read as A is the probability that the sample truly is A, $p(A)$, and that it is correctly read $(1 - d)$, plus the probability that the sample is not A, $1 - P(A)$ and that it is incorrectly read (d), or:

$$\begin{aligned} P(\text{read as A}) &= (1-d)P(A) + d(1-P(A)) \\ &= (1-2d)P(A) + d \end{aligned}$$

Thus under the simplest form of the PT+N model, the expected value of probability judgments is:

[1] It is worth noting that other non-sampling based approaches have been proposed to account for distortions in people's use of explicit probabilities in decision-making (e.g. Zhang & Maloney, 2012; Zhang et al., 2020). Further theorizing might extend these accounts to also describe the generation of probability estimates, so that a probabilistic account of beliefs might not rest entirely on the assumption of sampling from mental models.

$$E[\hat{P}_{PT+N}(A)] = (1 - 2d)P(A) + d$$

By assumption, a maximum of 50% of samples can be misread, so that d is a number in the range $[0, 1/2]$. The PT+N theory provides a unified account for a wide variety of biases in probability judgment that were previously attributed to different types of heuristics, as well as novel biases identified based on the model's predictions (Costello & Watts, 2014, 2016, 2017, 2018).

Meanwhile, Zhu, Sanborn, & Chater (2020) have proposed a Bayesian model of probability judgment they call the "Bayesian Sampler." Under this model, probability judgment is itself seen as a process of Bayesian inference. To judge the probability of an event, a limited number of samples are again drawn from a mental model of the event. Then, those "observed" samples are integrated with a prior over probabilities to produce a probability judgment. This prior takes the form of a symmetric Beta distribution, $Beta(\beta, \beta)$. After observing $S(A)$ successes and $N - S(A)$ failures, the posterior over probabilities is distributed $Beta(\beta + S(A), \beta + N - S(A))$. Zhu and colleagues (2020) assume that people report the mean of their posterior probability estimates. For any Beta distribution $x \sim Beta(a, b)$, $E[x] = \frac{a}{a+b}$. So, the expected probability estimate is a linear function of S, N, and β.

$$\hat{P}_{BS}(A) = \frac{S(A)}{N + 2\beta} + \frac{\beta}{N + 2\beta}$$

The expected value of the estimate can then be written in terms of the expected number of successes, or $P(A) \cdot N$. Under the simplest version of the Bayesian Sampler model, this gives the following formula:

$$E[\hat{P}_{BS}(A)] = \frac{N}{N + 2\beta}P(A) + \frac{\beta}{N + 2\beta}$$

Like the PT+N model, the Bayesian Sampler model accounts for a wide array of biases in probability judgments, including the novel biases identified by Costello and Watts (Costello & Watts, 2014, 2016). In fact, important equivalencies can be drawn between the two models. Zhu and colleagues (2020) show that the N and β parameters of their model can be related to the d parameter of the PT+N model via the following bridging formula:

$$d = \frac{\beta}{N + 2\beta}$$

Thus, in many cases the effect of a Bayesian prior is identical to the effect of noise in the PT+N model (at least in expectation). But, rather than merely perturbing people's probability judgments, this prior can be seen as regularizing these judgments away from extreme values. Zhu and colleagues (2020) argue that such regularization can be adaptive in cases where only a small number of mental samples can be drawn. For instance, consider someone estimating the probability that they can swim across a lake, outrun an animal, or win a hand of poker: if a mental simulation of these events produces two samples indicating success, one might conclude these are all certain victories and thereby be too willing to assume risk. A regularizing prior pushes these estimates away from extremes, thereby promoting better decision-making when mental samples are sparse. However, this hedging comes at the cost of systematic incoherence and biases.

Two accounts of conditional probability judgments. By explaining the incoherence of human probability judgments using coherent mental probabilities, both models have the potential to rescue the larger project of Bayesian cognitive science as applied to everyday beliefs (Chater et al., 2020). However, the two models diverge substantially in their treatment of conditional probability judgments. Bayesian cognitive theories are fundamentally theories of inductive reasoning: Bayes' rule describes how existing beliefs should be updated conditional on the observation of different kinds of evidence. So, treatment of the conditioning of beliefs is at the heart of these theories.

According to the Bayesian sampler model, conditioning is something that happens in the mental model of the events, not as part of the process of rendering probability judgments. By not assigning any special status to conditional probability judgments, the Bayesian Sampler theory fits neatly into the larger Bayesian project of cognitive science: probability judgments are simply another judgment process applied to the outputs of other (ideally Bayesian) mental models (Chater et al., 2020).

In contrast, the PT+N model presents a constructive account of conditional probability judgments that is fundamentally non-Bayesian (Costello & Watts, 2016). According to the PT+N model, conditional probabilities $P(A|B)$ are estimated by a two-stage sampling procedure: first both events A and B are sampled with noise, and then a second noisy process computes the ratio of the events read as A and B over events read as B. The PT+N model predicts conditional probability estimates using the following equation:

$$P_e(A|B) = \frac{(1-2d)^2 P(A \wedge B) + d(1-2d)\Big(P(A) + P(B)\Big) + d^2}{(1-2d)P(B) + d}$$

This non-Bayesian account of conditional probabiliity judgments separates the PT+N theory quite fundamentally from the Bayesian Sampler and the larger project of Bayesian cognitive science.

Comparing the models

Zhu, Sanborn, and Chater (2020) compared their Bayesian Sampler model against Costello & Watts' (2014; 2016; 2017; 2018) PT+N model as explanations for human probability judgments in two experiments. Unfortunately, their results were somewhat equivocal. They fit both "simple" and "complex" versions of each model, where the complex versions of these models introduce additional parameters d' and N' that allow for different patterns of judgments for conjunctive and disjunctive judgments as compared

with simple probability judgments. These additional parameters are crucial to both models' explanations of conjunction and disjunction fallacies—key findings in the probability judgment literature (Costello & Watts, 2017; Zhu et al., 2020). They compared the fits of these models to data via Bayesian Information Criteria score. Table 1 below presents the total BIC scores computed for each model as originally fit, using the authors' original code and saved model outputs (Zhu et al., 2020, Supplementary materials).

Table 1

Original model fitting results with best-fitting model in bold face.

Model	Exp. 1	Exp. 2
Bayesian Sampler simple	**956.9**	**-5371.9**
Bayesian Sampler complex	1174.4	-5099.3
PT+N simple	1257.5	-4901.1
PT+N complex	1039.7	-5159.6
Bayesian Sampler avg.	1065.6	-5235.6
PT+N avg.	1148.6	-5030.4

In both experiments, the simple version of the Bayesian Sampler scores best, but the complex version of the PT+N model comes in second (lower BIC scores are better). It is not obvious what conclusions should be drawn from these results. Of course, the simple Bayesian Sampler model appears to win by the numbers. But conjunction and disjunction fallacies are extremely robust empirical findings (Mellers et al., 2001; Sides et al., 2002) and are clearly present in the data collected by Zhu and colleagues (2020). So we might justifiably rule out the simple variants of the models on the grounds that they will fail to capture important qualitative features of the data, which could instead favor the PT+N theory. For their part, Zhu and colleagues (2020) chose to average the simple and complex model scores together. This approach somewhat favors the Bayesian Sampler theory

overall, though they are cautious about drawing strong conclusions in favor of their theory.

Predictive generalization and model complexity. A chief goal of scientific models is to make accurate predictions about future observations. When comparing and selecting models based on their predictive accuracy, it is crucial to consider how the models will perform on new, as-yet-unseen data. The traditional approach to this issue is to compute the fit of the model to the observed data, but to then correct for the potential for these models to "overfit" these data. Typically this correction comes in the form of a "complexity penalty," penalizing models in proportion to their flexibility for accommodating different patterns of data (Gelman, Hwang, & Vehtari, 2014). Zhu and colleagues (2020) warn that BIC, which penalizes models based solely on the number of parameters in the model, cannot fully account for the differences in the competing models' complexity (also see Piantadosi, 2018).

There are at least three challenges to accounting for model complexity in the comparison of the PT+N and Bayesian Sampler models. First, the models differ not only in the number of parameters but in the domain of those parameters. Zhu and colleagues (2020) assume that the Bayesian Sampler model's prior distribution should reflect ignorance or a lack of information. A uniform prior, Beta(1, 1) is the most obvious choice in this case, but theoretical arguments can also be made for Beta(0.5, 0.5) (Jeffrey's prior), Beta(0,0) (Haldene's prior), or perhaps other symmetric beta distributions Beta(β,β) with $\beta \in [0, 1]$ (Jaynes, 2003). Via the bridging conditions, they show that assuming $\beta \in [0, 1]$ restricts the "noise level" for the Bayesian Sampler, represented in implied d under the PT+N model, to fall within [0, 1/3] (approaching 1/3 as N approaches 1), whereas the PT+N model permits noise values in $[0, 1/2]$. d values in ther ange $[1/3, 1/2]$ imply β values greater than 1, corresponding to an informative prior over probabilities.

Second, it is not immediately clear how the models' structural differences impact their flexibility. Zhu and colleagues' (2020) bridging condition makes it clear that the PT+N model is more flexible when it comes to predicting unconditional probabilities. But,

what impact does the PT+N model's treatment of conditional probabilities have on model complexity? Is this component of the model a sort of Ptolemaic epicycle, adding complexity to the model that should be penalized? Or, does it constitute a commitment to novel predictions that thereby constrain its flexibility? Determining model complexity a priori is not always straightforward when the models being compared differ structurally.

Third, one potential explanation for the relative weakness of the "complex" model variants is Zhu and colleagues' (2020) use of "unpooled" models, where parameters are estimated independently for each individual participant. In contrast, a comparison of fully pooled variants of the simple and complex models (where a single population parameter is used for all participants) would require adding only one extra parameter to the penalty term. If there is limited heterogeneity across individuals, then adding a parameter for each participant may effectively over-penalize the complex variants relative to the simple variants. Partial pooling is a solution that balances between these extreme approaches, allowing for an accounting of heterogeneity without over-penalizing in cases where heterogeneity is low.

Finally, it is worth recognizing that formal measures of model complexity cannot be expected to perfectly track notions of simplicity or elegance in scientific explanation (for some related discussions, see Kuhn, 1977; Piantadosi, 2018; Sober, 2002). For instance, even if the PT+N model's account of conditional probability judgments constrains its flexibility empirically, it seems clear this added component makes it more complex as a putative scientific explanation.

The present work. Here, I cast both the Bayesian Sampler and PT+N models into a Bayesian data analysis framework that may permit a more decisive comparison. First, Bayesian data analysis allows issues of model complexity to be addressed through comparisons of model fit based newer information criteria, such as the widely-applicable information criterion (WAIC) and Pareto smoothed importance sampling approximate leave-one-out cross validation (PSIS-LOO). In particular, rather than estimating model fit

and then penalizing for model complexity, PSIS-LOO estimates out-of-sample prediction performance directly by estimating the expected log predictive density ($\widehat{\text{elpd}}$) of the model, or the expected probability of new unseen data (Gelman, Hwang, & Vehtari, 2014; Vehtari et al., 2017). From these calculations, an estimate of model complexity ($\hat{p}_{\text{LOO}}$) can also be derived.

In addition, the Bayesian framework supports straightforward implementation of hierarchical versions of these models with partial pooling. This allows for information about model parameters to be shared across participants, resulting in potential improvements to out-of-sample prediction, reductions in model complexity, and a more realistic test of the models.

Methods

Data selection

Zhu, Sanborn, & Chater (2020) conducted two experiments to compare the PT+N and Bayesian Sampler theories. These experiments asked participants to judge the probability of different events in various combinations. Following prior work by Costello and Watts (e.g. 2016, 2018), both experiments focused on the everyday events of different kinds of weather.

Experiment 1 asked about the events [icy, frosty] and [normal, typical] (e.g. "what is the probability that the weather in England is normal and not typical?"). The authors' goal was to ask about highly correlated events, but the events used are perhaps nearly perfectly correlated. Because the terms used to describe these events are nearly synonymous, there is a concern about the interpretation of the statements evaluated in this experiment. This is especially clear, as the authors note, for disjunctive query trials such as "normal or typical," where "or typical" might not be read as a disjunction but rather an elaborative clause. In light of these concerns, I excluded the disjunctive trials from

Experiment 1 from my analyses.

Experiment 2 focused on more moderately correlated events, [cold, rainy] and [windy, cloudy], that do not admit these misinterpretations. In addition, a third experimental condition asking about [warm, snowy] was also included in the experiment, but was dropped from the analyses reported in the paper. Exploring the raw responses from this condition reveals a substantial fraction of "zero" and "one" responses for certain trials. This may reflect a different response process than was intended. For instance, some participants may have engaged in deductive reasoning to judge that it is not possible for the weather to at once be warm and snowy, and therefore responded with zero—failing to properly consider that it is possible (at least logically) for it to be warm and snowy at different times within the same day. Given these potentially aberrant responses, I followed Zhu and colleagues (2020) in ignoring data from this condition.

Modeling

I implement several variants of the Bayesian Sampler and PT+N models in a Bayesian framework. These models were implemented in the probabilistic programming language Numpyro. All code and results are available as supplemental materials (https://github.com/derekpowell/bayesian-sampler).

Bayesian implementations of the models. The PT+N model defines expected probability judgments (P_e) as:

$$
\begin{aligned}
P_e(A) &= (1-2d)P(A) + d \\
P_e(A \wedge B) &= (1-d')P(A \wedge B) + d' \\
P_e(A \vee B) &= (1-d')P(A \vee B) + d' \\
P_e(A|B) &= \frac{(1-2d)^2 P(A \wedge B) + d(1-2d)\Big(P(A) + P(B)\Big) + d^2}{(1-2d)P(B) + d}
\end{aligned}
$$

In contrast, the Bayesian Sampler model defines expected probability judgments as:

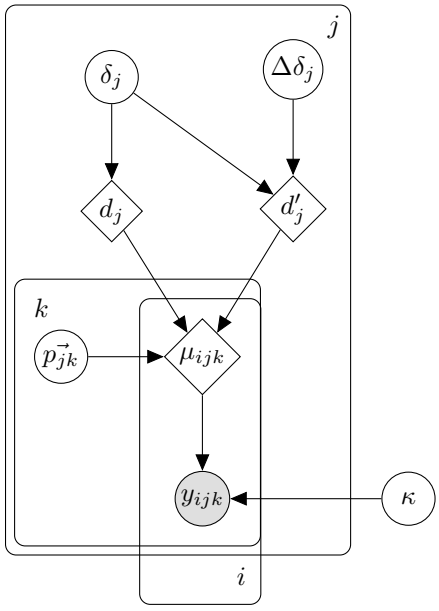
$$
\begin{aligned}
P_e(A) &= \frac{N}{N+2\beta}P(A) + \frac{\beta}{N+2\beta} \\
P_e(A \wedge B) &= \frac{N'}{N'+2\beta}P(A \wedge B) + \frac{\beta}{N'+2\beta} \\
P_e(A \vee B) &= \frac{N'}{N'+2\beta}P(A \vee B) + \frac{\beta}{N'+2\beta} \\
P_e(A|B) &= \frac{N}{N+2\beta}P(A|B) + \frac{\beta}{N+2\beta}
\end{aligned}
$$

Fixing d and d' or N and N' equal yields the "simple" variant of each of the models, which treat conjunctive and disjunctive probability judgments identically to simple probability judgments.

Notice that for each model the probability judgments depend on underlying subjective probabilities, derived from a mental sampling process. These subjective probabilities are unobserved, and must be estimated as a latent variable. Here, they are represented with a four-dimensional dirichlet distribution for each subject, representing the probability of the elementary events ($A \wedge B$, $\neg A \wedge B$, $A \wedge \neg B$, $\neg A \wedge \neg B$).

Zhu, Sanborn & Chater (2020) implement completely unpooled models with separate d, d', N, N', and β parameters for each participant. Although hierarchical models with partial pooling might be expected to better account for the data and offer a better test of the models, for consistency and comparison with Zhu et al.'s (2020) analyses, I first estimated implementations of these unpooled models. Figure 1 displays the translation of the PT+N model into the Bayesian framework, along with a plate diagram representing the dependencies among parameters.

The function f_{PT+N} computes the expected probability estimate using the underlying subjective probability p, the noise parameters d and d', and the relevant equation as defined by the PT+N theory (see supplemental materials for implementation details).



$$
\begin{aligned}
y_{ijk} &\sim Beta(\mu_{ijk}\kappa, (1-\mu_{ijk})\kappa) \\
\mu_{ijk} &= f_{PT+N}(\overrightarrow{p_{jk}}, d_j, d_j) \\
d_j &= \frac{1}{2}\ \text{logistic}(\delta_j) \\
d'_j &= \frac{1}{2}\ \text{logistic}(\delta_j + \exp(\Delta\delta_j)) \\
\overrightarrow{p_{jk}} &\sim Dirichlet(\overrightarrow{1}) \\
\delta_j &\sim Normal(0,1) \\
\Delta\delta_j &\sim Normal(0,1) \\
\kappa &\sim \textit{Half-Cauchy}(20)
\end{aligned}
$$

Figure 1. Complex unpooled PT+N model diagram and formula specifications. Circular nodes are parameters, shaded nodes are observations, and squared nodes are deterministic functions of parameters. Plates signify values defined for i trials, j participants, and k conditions.

Prior predictive checks were conducted for all models to select priors that would be uninformative or minimally informative on the scale of the model parameters d and d'.

Recall that Zhu and colleagues (2020) identified a bridging condition relating β and N in the Bayesian Sampler model to the d parameter of the PT+N model. To support direct comparisons of the models, I parameterize the Bayesian Sampler model according to the implied d and d', rather than directly according to its β, N, and N' parameters.[2] I

[2] Strictly speaking, under the original form of the Bayesian sampler model, N and N' are discrete parameters representing the number of distinct independent samples drawn. Given a particular implied d, this could create constraints on the possible values of d', assuming β is held constant. However, Zhu and colleagues (2020) also consider the possibility that people draw non-independent mental samples, in which case N and N' would represent the *effective number of samples*, accounting for their autocorrelation. In this case, we could treat this effective number of samples as a continuous quantity, and therefore imagine there are no clear constraints on d and d' except the stipulation that $d \leq d'$.

constrain d to $[0, 1/3]$ for the Bayesian Sampler model to reflect the assumption that $\beta \in [0, 1]$. This allows the same priors to be used for the corresponding Bayesian Sampler and PT+N models, simplifying their comparison.

The Bayesian Sampler model is therefore identical to the PT+N model save for the changes to μ_{ijk}, d, and d' shown below:

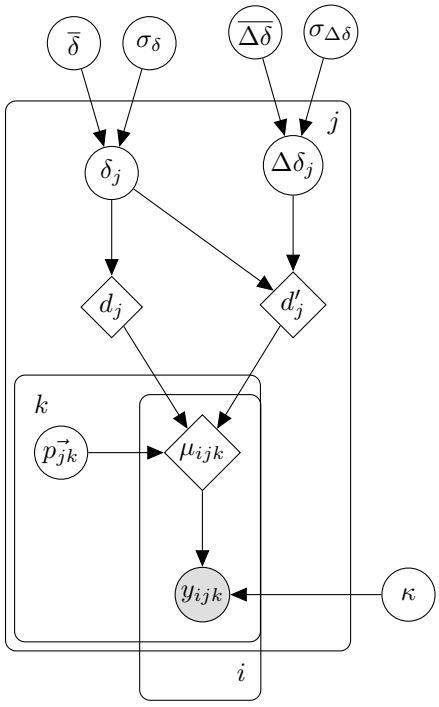
$$\begin{aligned} \mu_{ijk} &= f_{BS}(\overrightarrow{p_{jk}}, d_j, d_j) \\ d_j &= \frac{1}{3}\ \text{logistic}(\delta_j) \\ d'_j &= \frac{1}{3}\ \text{logistic}\Big(\delta_j + \exp(\Delta\delta_j)\Big) \end{aligned}$$

Where the function f_{BS} computes the expected probability estimate as prescribed by the Bayesian Sampler theory.

Hierarchical implementations of the models. Both of these models can also be implemented as hierarchical models with partial pooling for the d and d' parameters (implicitly, for N and N' in the case of the Bayesian Sampler). This partial pooling can help to regularize parameter estimates and improve out-of-sample predictive performance. In addition, partial pooling effectively reduces model complexity, and could support more realistic comparison between the "simple" and "complex" variants of the models.

The hierarchical implementation adds parameters for the population-level d and d' as well as a parameter controlling the standard deviation of the distribution for the subject-level effects. For ease of interpretation, the centered parameterization is shown below, although the actual models used a non-centered parameterization to improve sampling efficiency (Papaspiliopoulos et al., 2007). Figure 2 displays the translation of a hierarchical implementation of the Bayesian Sampler model into the Bayesian framework, along with a plate diagram representing the dependencies among parameters.

Finally, I also explored fitting a hierarchical version of the Bayesian Sampler model



$$
\begin{aligned}
y_{ijk} &\sim Beta(\mu_{ijk}\kappa, (1-\mu_{ijk})\kappa) \\
\mu_{ijk} &= f_{BS}(\overrightarrow{p_{jk}}, d_j, d'_j) \\
d_j &= \frac{1}{3}\ \text{logistic}(\delta_j) \\
d'_j &= \frac{1}{3}\ \text{logistic}(\delta_j + \exp(\Delta\delta_j)) \\
\overrightarrow{p_{jk}} &\sim Dirichlet(\overrightarrow{1}) \\
\bar{\delta} &\sim Normal(-1, 1) \\
\overline{\Delta\delta} &\sim Normal(0, .50) \\
log(\sigma_\delta) &\sim Normal(-1, 1) \\
log(\sigma_{\Delta\delta}) &\sim Normal(-1, 1) \\
\delta_j &\sim Normal(\bar{\delta}, \sigma_\delta) \\
\Delta\delta_j &\sim Normal(\overline{\Delta\delta}, \sigma_{\Delta\delta}) \\
\kappa &\sim \textit{Half-Cauchy}(20)
\end{aligned}
$$

Figure 2. Hierarchical complex Bayesian Sampler model diagram and formula specifications. Circular nodes are parameters, shaded nodes are observations, and squared nodes are deterministic functions of parameters. Plates signify values defined for i trials, j participants, and k conditions.

that allowed values of $\beta > 1$. Restricting β to [0,1] restricts the prior distribution of the Bayesian sampler to the class of "ignorance priors" (Zhu et al., 2020). However, it is also possible that people bring informative priors to the probability judgment task. Indeed, Zhu and colleagues (2020) acknowledge there are situations where an informative prior may be warranted (see e.g., Fennell & Baddeley, 2012). If β is unrestricted, allowed to fall in the domain $[0, \infty]$ then the Bayesian Sampler model becomes more flexible, allowing for equivalent "noise" levels in the same $[0, 1/2]$ range as the PT+N model. That is, through the bridging condition, the implied d approaches 1/2 in the limit as $N \to 1$ and $\beta \to \infty$. Though it would seem a more fundamental change, this same model may also be seen as a version of the PT+N theory that jettisons its constructive account of conditional probability judgment. Thus, fitting this additional unrestricted model allows for a complete comparison of the models along both of their differing dimensions.

Results

I fit each of the models specified above to data from Zhu et al's (2020) Experiment 1 and 2 and estimated the expected log predictive density with PSIS-LOO ($\widehat{\text{elpd}}_{\text{LOO}}$) for each combination. Compared with BIC, $\widehat{\text{elpd}}_{\text{LOO}}$ offers a more sophisticated account of model complexity and is more appropriate in the "$\mathcal{M}$-open" case; situations where we do not know if any of the models being compared are the "true" model (Vehtari et al., 2019). Model posteriors were estimated using the Numpyro (Phan et al., 2019) implementation of the No-U-Turn Hamiltonian Markov chain Monte Carlo (MCMC) sampler. For each model, four MCMC chains of 2000 iterations were sampled after 2000 iterations of warmup and all passed convergence tests according to $\hat{R}$ (see Gelman, Carlin, et al., 2014). Figure 3 below displays the estimated differences in $\widehat{\text{elpd}}_{\text{LOO}}$ scores for each of the models as compared to the best-scoring model.

Data from Experiment 1 favor "complex" variants of the Bayesian Sampler model compared with the "simple" variants and all versions of the PT+N model. However, there

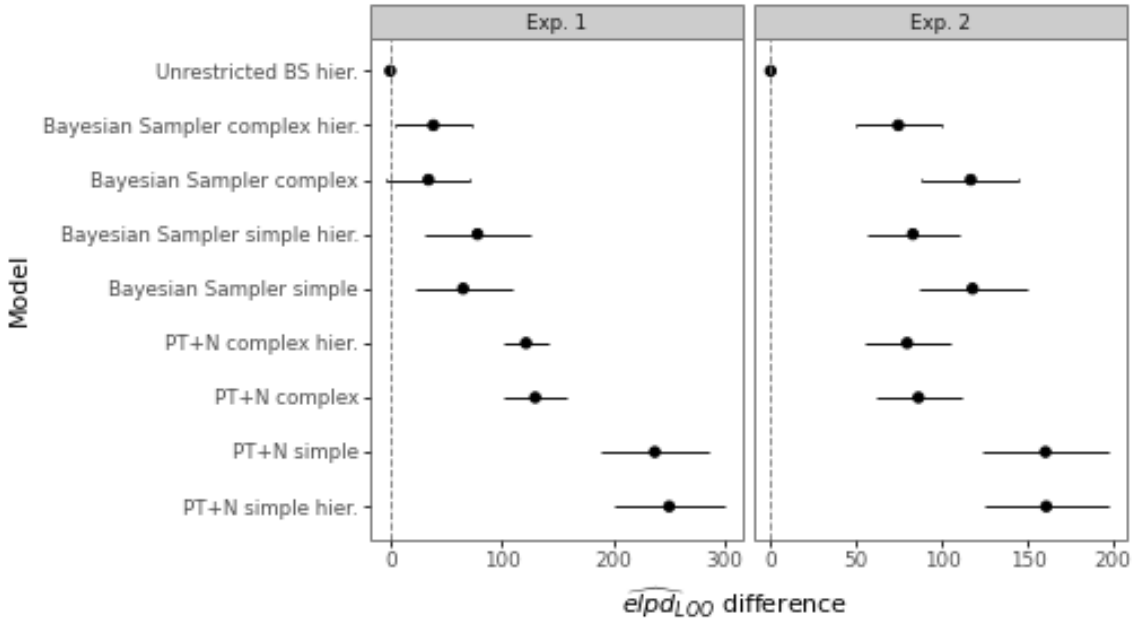


Figure 3. Model comparison results for data from Experiments 1 and 2. Error bars indicate two standard errors of the estimates. Typically, a difference of greater than two standard errors is taken as clear evidence for the superiority of the lower-scoring model (Sivula et al., 2020).

is no clear single winner. As shown in Figure 3, the best-scoring model is an unrestricted variant of the Bayesian Sampler allows for people to bring informative priors to the probability judgment task, that is, a Bayesian Sampler model allowing $\beta \in [0, \infty]$ (greater values of $\widehat{\text{elpd}}_{\text{LOO}}$ are better). However, the complex and complex hierarchical implementations of the Bayesian Sampler assuming uninformative priors (i.e., restricting $\beta \in [0, 1]$) have $\widehat{\text{elpd}}_{\text{LOO}}$ scores within two standard errors of the difference, indicating that these models are also plausible (Sivula et al., 2020).

Data from Experiment 2 more decisively reveal a single winning model: the hierarchical "unrestricted" implementation of the Bayesian Sampler model allowing for informative priors.

Figure 4 shows the posterior distributions of the population-level d and d' parameters

inferred from the unrestricted Bayesian Sampler model. In Experiment 2, population-level estimates of d' are greater than 1/3, as are a substantial number of participant-level estimates for d (37 of 83). These values fall outside the range implied by the assumption of "ignorance priors" in the Bayesian Sampler model. Parameters fit to the data from Experiment 1 are more consistent with this assumption, although a substantial proportion of individual participants' d and d' estimates also lie outside this range (11 of 59 for d, 18 of 59 for d'). The finding that there are clear differences in d and d' estimated across experiments suggest that the mental sampling processes producing estimates vary in the different conditions, either in terms of the number of samples that are drawn, the noise in reading those samples, or the form of the prior distribution assumed by participants in each context.

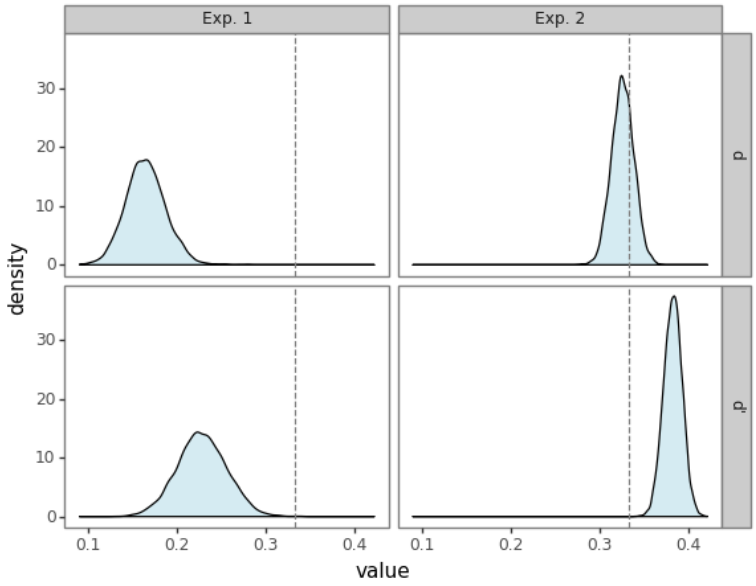


Figure 4. Posterior density of population-level d and d' parameters estimated from the unrestricted hierarchical Bayesian Sampler model for data from Experiments 1 and 2. Dashed line indicates theoretical maximum values for Bayesian Sampler model with uninformative priors.

Recall, the unrestricted Bayesian Sampler model may also be seen as a version of the

PT+N theory that excises its constructive account of conditional probability judgments. So, comparing the PT+N models to the unrestricted Bayesian Sampler model offers a test of the PT+N's constructive account of conditional probability jugments. Comparing predictions from the unrestricted Bayesian Sampler model and the best-fitting PT+N model, we see that the Bayesian Sampler model better captures these judgments from both experiments (see Table 2). This is true for both query-level averages as well as for individual participants' trial responses. And comparing their predictions specifically for conditional probability judgments, the unrestricted Bayesian Sampler again provides a better fit, especially for the modestly-correlated events of Experiment 2 (Exp 1: response-level $r = .91$ vs. .89 ($p < .05$); query-level $r = .98$ vs. .97; Exp 2: response-level $r = .72$ vs. .67 ($p < .01$); query-level $r = .92$ vs. .85) These findings suggest that conditioning is better seen as part of a mental model of the events than as part of the probability judgment process.

Table 2 also presents estimates of $\hat{p}_{\text{LOO}}$, the estimate of the effective number of parameters for each model. Compared to the Bayesian Sampler model, the PT+N model with its constructive account of conditional probability judgments has a similar penalty term estimate when fit to Experiment 1, but has a smaller penalty term when fit to Experiment 2, despite having the same parameterization in terms of d and d'. Although its special treatment of conditional probability judgments makes it more complex as a putative scientific explanation, this structural component appears to actually constrain its predictive flexibility. However, this constraint leads to a worse-fitting model.

Perhaps surprisingly, the Bayesian Sampler with unrestricted β actually receives a smaller penalty term than the more "restricted" version of the model for the data in Experiment 2. This is at first counterintuitive. However, model complexity depends not only on the model and priors, but also the observed data (see Gelman, Hwang, & Vehtari, 2014). To illustrate, Gelman and colleagues consider a case where a parameter is constrained to be positive and its value is then estimated from data (2014). If the

estimated value is some very large positive number, then the constraint won't have been very informative. But, if the estimated value is very close to zero, then the constraint that the parameter is positive will provide substantial information and the model's penalty term will therefore be smaller. Here, it seems reasonable to conjecture that because the implied d and d' estimated from Experiment 2's data are both very near 1/3 under this model, the restriction results in posterior estimates of the linear parameters that are relatively far from the prior, which can result in a greater penalty.

The dependence of complexity penalties on observed data may strike some as an undesirable feature of model comparison through information criteria like PSIS-LOO. Indeed, it is worth acknowledging that principled model comparison is still an area of active inquiry, with differing perspectives (e.g. Gronau & Wagenmakers, 2019; Vehtari et al., 2019). Fortunately, in this case, comparisons between the models do not rest solely on differences in estimated model complexity.

Table 2

Bayesian model comparison results with best scoring model in bold face.

	Experiment 1				Experiment 2			
Model	$\widehat{\text{elpd}}_{\text{LOO}}$	$\hat{p}_{\text{LOO}}$	r_{resp}	r_{query}	$\widehat{\text{elpd}}_{\text{LOO}}$	$\hat{p}_{\text{LOO}}$	r_{resp}	r_{query}
Unrestricted BS hier.	**1118.7**	**259.3**	**0.884**	**0.964**	**1978.6**	**366.4**	**0.688**	**0.878**
Bayesian Sampler complex hier.	1088.3	269.6	0.883	0.960	1912.3	395.0	0.675	0.852
Bayesian Sampler complex	1087.8	264.6	0.883	0.961	1876.9	443.4	0.679	0.848
Bayesian Sampler simple	1045.6	253.8	0.874	0.952	1861.2	419.1	0.667	0.831
Bayesian Sampler simple hier.	1039.6	259.1	0.874	0.953	1900.7	377.8	0.667	0.839
PT+N complex hier.	993.0	268.9	0.867	0.946	1902.4	351.5	0.658	0.835
PT+N complex	966.1	259.8	0.864	0.941	1886.9	395.5	0.667	0.840
PT+N simple hier.	864.0	250.0	0.839	0.919	1821.9	305.3	0.617	0.772
PT+N simple	863.5	245.5	0.838	0.918	1821.5	319.5	0.619	0.777
Relative Freq.	649.3	289.4	0.820	0.875	643.8	424.6	0.516	0.639

Finally, it is worth noting that the best of these models provide quite strong overall fits to the data, not just for the query averages, but also for the query averages across individual participants as seen from the correlations between predicted and observed responses in Table 2.

Discussion

By a fair margin, the models best accounting for the experimental data from Zhu and colleagues (2020) were implementations of the Bayesian Sampler theory, with the best single model being a version of the Bayesian Sampler model without restriction on the range of its β parameters. Alternatively, this model can also be seen as a variant of the PT+N model that removes its account of conditional probability judgments. Thus, what these findings indicate most clearly is that the Bayesian Sampler theory provides a superior account of conditional probability judgments in this task. In keeping with the larger theoretical framework of Bayesian cognitive science, the Bayesian Sampler theory assumes that subjective probabilities underlie people's probability judgments, and that conditional probability judgments are produced by Bayesian conditioning occurring in their mental models of the events in question, rather than as arising from the probability judgment process (Chater et al., 2020; Zhu et al., 2020).

Zhu and colleagues' experiments cast some doubt on their proposal that the priors of the Bayesian Sampler model should reflect "ignorance priors," symmetric Beta distributions with $\beta \in [0, 1]$. As a generic prior that would be used across contexts, this class of uninformative priors has clear appeal. Nevertheless, many participants' estimated d and d' parameters fell outside the $[0, 1/3]$ range implied by "ignorance" priors. Given the different values of d and d' across Experiments 1 and 2, it seems likely that people bring informative priors to at least some probability judgment tasks.

How might these theories explain the difference in implied d parameters between

Experiments 1 and 2? Under the Bayesian Sampler theory, it could be that people bring informative and domain-specific priors to the probability judgment task, so that participants' priors differed across the different contexts of Experiments 1 and 2. However, these experimental contexts are highly similar, so it is not entirely clear why this should be the case. Alternatively, it could be that participants held identical informative priors in both contexts, but that they were able to more effectively draw mental samples for the queries presented in Experiment 1 than in Experiment 2 (resulting in lower d). Similarly, under a noise-based account, it could be that the mental samples are read more noisily in Experiment 2 than Experiment 1. Unfortunately, teasing apart these different accounts is likely to prove challenging. First, the same sort of factors that could be expected to influence the number of samples could also plausibly influence "noise" (e.g. task complexity, expertise). Moreover, pinning down specific components of the Bayesian Sampler model is a challenge, as the β and N parameters are not uniquely identifiable from this kind of judgment data alone.

In addition, the implied d and d' noise parameters also varied across individuals. Of note, some participants' implied d and d' parameters *were* consistent with the class of ignorance priors, even in Experiment 2. Further research should explore this heterogeneity within and across individuals.

One thing this model comparison has not decided, and likely *cannot* decide, is whether distortions of probability judgments are products of mental noise or of further reasoning processes. Given the two models' tight connections via the bridging condition (Zhu et al., 2020), it may not be possible to draw decisive conclusions here. Moreover, the theories may not be in any real competition over this point: Zhu and colleagues consider that "noise" might give an algorithmic-level solution to the computation-level task defined by the Bayesian Sampler (2020).

Finally, some of the most interesting implications of these models go well beyond the

probability judgment task itself: the models both support a probabilistic account of beliefs (Chater et al., 2020). Indeed, by representing the "true" subjective probabilities as a latent variable, Bayesian data analysis allows those underlying credences to be inferred. Examining the model posteriors here reveals these estimates often come with considerable uncertainty, but at least for some participants they can be estimated with useful levels of precision. Of course, Zhu and colleagues' (2020) experiments were never designed for this purpose. Future research could explore how estimates of people's credences might be made more reliable, and how inferences about these mental probabilities might be integrated with other Bayesian models of reasoning (e.g. Franke et al., 2016; Griffiths & Tenenbaum, 2006; Jern et al., 2014). For instance, people's responses in various reasoning tasks are often explicitly related to inferred subjective mental probabilities (e.g. Jern et al. (2014)), so accounting for biases in those reports may permit more rigorous model testing. One particularly promising direction could be to integrate these models with formal models of belief revision, which might then shed new light on these fundamental cognitive processes (e.g. Cook & Lewandowsky, 2016; Jern et al., 2014; Powell et al., 2018).

References

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. (2020). Probabilistic Biases Meet the Bayesian Brain. *Current Directions in Psychological Science*, *29*(5), 506–512. https://doi.org/10.1177/0963721420954801

Cook, J., & Lewandowsky, S. (2016). Rational Irrationality: Modeling Climate Change Belief Polarization Using Bayesian Networks. *Topics in Cognitive Science*, *8*(1), 160–179. https://doi.org/10.1111/tops.12186

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480. https://doi.org/10.1037/a0037010

Costello, F., & Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, *89*, 106–133. https://doi.org/10.1016/j.cogpsych.2016.06.006

Costello, F., & Watts, P. (2017). Explaining High Conjunction Fallacy Rates: The Probability Theory Plus Noise Account. *Journal of Behavioral Decision Making*, *30*(2), 304–321. https://doi.org/10.1002/bdm.1936 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.1936.

Costello, F., & Watts, P. (2018). Invariants in probabilistic reasoning. *Cognitive Psychology*, *100*, 1–16. https://doi.org/10.1016/j.cogpsych.2017.11.003

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1–25. https://doi.org/10.1016/j.cogpsych.2017.05.001

Fennell, J., & Baddeley, R. (2012). Uncertainty plus prior equals rational bias: An intuitive Bayesian probability weighting function. *Psychological Review*, *119*(4), 878–887. https://doi.org/10.1037/a0029346

Franke, M., Dablander, F., Scholler, A., Bennett, E., Degen, J., Tessler, M. H., Kao, J., & Goodman, N. D. (2016). What does the crowd believe? A hierarchical approach to estimating subjective beliefs from empirical data, 6.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third edition). CRC Press.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal Predictions in Everyday Cognition. *Psychological Science*, *17*(9), 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x

Gronau, Q. F., & Wagenmakers, E.-J. (2019). Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection. *Computational Brain & Behavior*, *2*(1), 1–11. https://doi.org/10.1007/s42113-018-0011-7

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science* (G. L. Bretthorst, Ed.). Cambridge University Press.

Jern, A., Chang, K.-m. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224. https://doi.org/10.1037/a0035941

Kahneman, D. (2013). *Thinking, Fast and Slow* (1st edition). Farrar, Straus and Giroux.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object Perception as Bayesian Inference. *Annual Review of Psychology*, *55*(1), 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. University of Chicago press.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*(3), 617–648. https://doi.org/10.1037/a0028719

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do Frequency Representations Eliminate Conjunction Effects? An Exercise in Adversarial Collaboration. *Psychological Science*, *12*(4), 269–275. https://doi.org/10.1111/1467-9280.00350

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A General Framework for the Parametrization of Hierarchical Models. *Statistical Science*, *22*(1). https://doi.org/10.1214/088342307000000014

Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv:1912.11554 [cs, stat]*.

Piantadosi, S. T. (2018). One parameter is always enough. *AIP Advances*, *8*(9), 095118. https://doi.org/10.1063/1.5031956

Powell, D., Weisman, K., & Markman, E. M. (2018). Articulating lay theories through graphical models: A study of beliefs surrounding vaccination decisions, 6.

Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, *30*(2), 191–198. https://doi.org/10.3758/BF03195280

Sivula, T., Magnusson, M., & Vehtari, A. (2020). Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison. *arXiv:2008.10296 [stat]*.

Sober, E. (2002). What is the problem of simplicity? In A. Zellner, H. A. Keuzenkamp, & M. McAleer (Eds.), *Simplicity, Inference and Modelling* (First, pp. 13–31). Cambridge University Press. https://doi.org/10.1017/CBO9780511493164.002

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Tversky, A., & Kahneman, D. (1983). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, *90*(4), 23.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567.

https://doi.org/http://dx.doi.org.ezproxy1.lib.asu.edu/10.1037/0033-295X.101.4.547 (US).

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Simpson, D. P., Yao, Y., & Gelman, A. (2019). Limitations of "Limitations of Bayesian Leave-one-out Cross-Validation for Model Selection". *Computational Brain & Behavior*, *2*(1), 22–27. https://doi.org/10.1007/s42113-018-0020-6

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245

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*. https://doi.org/10.3389/fnins.2012.00001

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

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*, *127*(5), 719–748. https://doi.org/10.1037/rev0000190