Computational Brain & Behavior (2023) 6:228–245
https://doi.org/10.1007/s42113-022-00164-z

ORIGINAL PAPER

Comparing Probabilistic Accounts of Probability Judgments

Derek Powell[1]

Accepted: 15 December 2022 / Published online: 14 February 2023
© Society for Mathematical Psychology 2023

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" (PT+N; Costello and Watts *Psychological Review*, *121*, 463–480, 2014) and "Bayesian Sampler" (Zhu et al. *Psychological Review*, *127*, 719–748, 2020) theories of probability judgments, both seek to account for these biases while maintaining that mental credences are fundamentally probabilistic. These models differ in their averaged predictions about people's conditional probability judgments and in their distributional predictions about their overall patterns of judgments. In particular, the Bayesian Sampler's Bayesian adjustment process predicts a truncated range of responses as well as a correlation between the average degree of bias and variability trial-to-trial. However, exploring these distributional predictions with participants' raw responses requires a careful treatment of rounding errors and exogenous response processes. Here, I cast these theories into a Bayesian data analysis framework that supports the treatment of these issues along with principled model comparison using information criteria. Comparing the fits of both models on data collected by (Zhu et al. *Psychological Review, 127*(5), 719–748 2020), I find these data are best explained by an account of biases based on "noise" in the sample-reading process but in which conditional probability judgments are produced by a process of conditioning in the mental model of the events, rather than in a two-stage mental sampling process as proposed by the PT+N model.

Keywords Probability judgments · Bayesian cognitive science · Heuristics and biases

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.

✉ Derek Powell
dmpowell@asu.edu

[1] School of Social and Behavioral Sciences, Arizona State University, Glendale, AZ 85306, USA

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 & 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 end products are probability judgments reflecting probabilistic credences perturbed by noise. In their model, the probability that a mental sample for an event A is correctly 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:

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

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

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

By assumption, a maximum of 50% of samples can be misread on average, so d is a number in the range $[0, 1/2]$. The overall consequence of the sample-reading noise will be to shrink probability estimates toward .50 in proportion to d. 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). For example, the PT+N theory offers an explanation for many instances of "conservatism" (Costello & Watts, 2014)—people's tendency to shy away from extreme probability judgments near 0 and 1, even when strong evidence warrants such judgments (e.g. Edwards, 1968, Erev et al., 1994).

Meanwhile, Zhu et al. (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 et al. (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.

$$\hat{P}_{BS}(A) = \frac{S(A)}{N+2\beta} + \frac{\beta}{N+2\beta} \tag{3}$$

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:

$$E[\hat{P}_{BS}(A)] = \frac{N}{N+2\beta} P(A) + \frac{\beta}{N+2\beta} \tag{4}$$

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 and Watts, 2014, 2016). In fact, important equivalencies can be drawn between the two models. Zhu et al. (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:

$$d = \frac{\beta}{N+2\beta} \tag{5}$$

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). A caveat to this is that the Bayesian Sampler theory restricts the parameterization of the equivalent d parameter compared to the PT+N model. Whereas PT+N assumes $d \in [0, 1/2]$, the Bayesian Sampler theory assumes an uninformatative prior parameter $\beta \in [0, 1]$, which in turn restricts the equivalent parameterization of $d \in [0, 1/3]$. But beyond this subtle difference of parameterization, there is a larger difference in interpretation: rather than merely perturbing people's probability judgments, this prior can be

[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.

seen as regularizing these judgments away from extreme values. Zhu et al. (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.

Differentiating Between the Models

The model's predictions can be distinguished on two levels: First, the models have distinct accounts of conditional probability judgments that make different predictions in terms of expected values. Second, the models present different process-level accounts of probability judgment that entail different predictions about the shape of the distribution of responses across trials.

Different 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 project of Bayesian 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. Schematically, the estimated probability can be written as:

$$\begin{aligned} P_e(A|B) &= P(\textit{read as } A|\textit{read as } B) \\ &= P(\textit{read as } A|B)P(B|\textit{read as } B) \\ &\quad + P(\textit{read as } A|\neg B)P(\neg B|\textit{read as } B) \end{aligned} \tag{6}$$

Substituting terms according to the PT+N model and then simplifying, the PT+N model predicts conditional probability estimates using the following equation:

$$P_e(A|B) = \frac{(1-2d)^2 P(A \wedge B) + d(1-2d)\big(P(A)+P(B)\big) + d^2}{(1-2d)P(B) + d} \tag{7}$$

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

Different Process-Level Accounts and Predicted Response Distributions

Although the model's predictions for unconditional probability judgments are identical in expectation (as seen via the bridging condition), the models posit different psychological processes underlying those judgments: sample reading noise in the PT+N model and Bayesian inference in the Bayesian Sampler model. These process-level differences imply different predictions about the distributions of people's judgments.

The models make qualitatively different distributional predictions on two fronts. First, the Bayesian Sampler predicts a clear relationship between the degree to which responses are shrunk toward .50 and the trial-by-trial variability in those responses. In both models, the amount of variability in trial-level responses is related to the number of mental samples drawn, N. In the Bayesian Sampler model, assuming β is relatively small, N should also help to determine the degree to which responses are shrunk toward .50. In contrast, in the PT+N model the variance across responses and degree of shrinkage are reasonably considered to be independent. Second, because the Bayesian Sampler describes a process of adjustment after the sampling process, in which people report the mean of their mental posterior over probabilities, the model also predicts a truncation of the response distribution in proportion to β and N (Chater et al., 2020; Sundh et al., 2021). That is, even when zero positive or negative samples are drawn, the mean of the posterior is drawn away from extreme responses of zero and one.

Modeling the distributions of raw responses holds clear promise for disentangling the models. However, there are at least three challenges to directly modeling raw human response data. First, both models are, strictly speaking,

discrete and so make a limited set of discrete predictions while assigning zero probability to responses outside that set. Second, and similarly, the truncation in the Bayesian Sampler model also assigns zero probability to responses beyond the truncated range. And third, from a cursory glance it is clear that a majority of human responses are rounded by some unknown degree, with most seemingly rounded to the nearest 5 or 10%. Given the combination of these factors, if fit directly to raw human data the posterior probability of both models is likely to be zero. I return to these challenges and my approaches to addressing them in the results.

Prior Comparisons of the Models

Comparison of Participant-Level Query-Averaged Responses

Zhu et al. (2020) compared their Bayesian Sampler model against Costello and Watts' (2014, 2016, 2017, 2018) PT+N model as explanations for human probability judgments in two experiments. Unfortunately, their results were somewhat equivocal.

Zhu et al. (2020) measured participants' judgments for each query (e.g. "what is the probability that it will be rainy") on three repeated trials. Their primary quantitative analysis fit the models separately to participants' average response to each query (averaged over three trials). These analyses compare human responses to the models' predictions *in expectation*. After fitting, Bayesian Information Criteria (BIC) values were computed for each participant, which were then used to approximate the posterior probability of each model for each participant, assuming a uniform prior. The researchers found that a preponderance of participants' responses were best-captured by the Bayesian Sampler model. However, a substantial number of participants were instead more strongly fit by the PT+N model.

Given that these models are proposing quite basic psychological processes, we might expect the same process to be shared across all people. But, the authors do not report on the overall posterior probability of each model if one model is assumed to explain all participants' responses. Such a comparison with these methods would likely be limited in a few ways. First, as they note (Zhu et al., 2020), BIC cannot fully account for the differences in the competing models' complexity (also see Piantadosi, 2018). Further, their "unpooled" analysis likely exaggerates the complexity of the models overall and may therefore affect comparisons between them. In contrast, hierarchical models with partial pooling offer a solution that balances between ignoring individual variation and allowing all parameters to vary freely, allowing for an accounting of heterogeneity without over-penalizing in cases where heterogeneity is low.

Comparison of Distributional Model Predictions

Rather than computing query-level averages across trials for each participant, examining the models' distributional predictions requires modeling participants' raw trial-by-trial responses. As mentioned above, this presents substantial challenges. To address these, Zhu et al. (2020) estimated discrete versions of the Bayesian Sampler and PT+N models by minimizing the Wasserstein distance between participants' raw responses and model predictions. The use of Wasserstein distance rather than a proper likelihood-based measure of model fit helped to minimize issues created by rounding and out-of-support responses, which could otherwise lead both models to assign probability zero to many observations.

Still, the results of this analysis were largely inconclusive with respect to differentiating the models (Zhu et al., 2020). Specifically, the quality of the fit for each model depended heavily on the maximum number of samples that is assumed possible. For small numbers of samples, the Bayesian Sampler model is clearly superior, but for larger numbers of samples, the PT+N model was found to better fit the data. Presumably, this is because with small numbers of samples the PT+N model is extremely constrained in the distinct discrete responses it can predict. For small N, both models predict only a limited set of distinct values are possible. However, whereas the size of that set is the same for each model, in the Bayesian Sampler the continuous β parameter can shift exactly what those discrete values are, providing it much greater flexibility. Yet, this additional model flexibility goes unpunished in comparisons based on Wasserstein distance.

In later work, Sundh et al. (2021) examined the distributional properties of participants' responses using indirect means, by regressing the variance of participants' responses across trials on their mean response for each query type. Their findings suggest that earlier fits with Wasserstein distance may have produced biased results (Sundh et al., 2021). They reported evidence for the truncation of responses and a correlation between variance and shrinkage parameter estimates across participants. However, their analysis did not enforce that the underlying probabilities driving participants' judgments be coherent (simply estimating the true probability as the mean across trials), nor did they evaluate how frequently participants gave out-of-support judgments that would be inconsistent with the Bayesian Sampler theory.

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. Two sets of analyses compared different aspects of the models' predictions: First, the model's were examined based on their predictions in expectation. This analysis allows for a test of their different accounts of conditional probability judgments and their parameterizations (i.e. the restriction of d to $[0, 1/2]$ versus $[0, 1/3]$). To preview the results, these analyses revealed the Bayesian Sampler's account of probability judgments to be superior, but could not distinguish between the different psychological processes proposed by the two models. A second set of analyses examined the model's distributional predictions to test their process-level accounts of probability judgments. As will be described, modeling response distributions directly presents a number of challenges, and so this second set of analyses required some minor additions and modifications to the models to permit their fitting to experimental data.

Both sets of analyses are supported by the use of a Bayesian Framework. First, Bayesian analyses allow issues of model complexity to be addressed through comparisons of model fit based on modern information criteria, such as Pareto smoothed importance sampling approximate leave-one-out cross validation (PSIS-LOO; Gelman et al. 2014, Vehtari et al. 2017).[2] Second, the Bayesian framework supports straightforward implementation of hierarchical versions of these models allowing 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. Finally, a Bayesian framework also supports new extensions of these models to directly model participants trial-level responses while accounting for rounding and out-of-distribution response errors. These extensions allow for principled probabilistic tests of the distributional predictions of the models.

[2]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 et al., 2014; Vehtari et al., 2017). From these calculations, an estimate of model complexity $\hat{p}_{\text{LOO}}$ can also be derived. However, it is worth recognizing that formal measures of model complexity will not always track notions of simplicity or elegance in scientific explanation (for some related discussions, see (Kuhn, 1977; Piantadosi, 2018; Sober, 2002)

Methods

Data Selection

Zhu et al. (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 et al. (2020) in ignoring data from this condition.

Modeling Results: Participant-Level Query-Averaged Responses

This first set of analyses compares the models' ability to capture participant's probability judgments in expectation, averaged over the three blocks on which they made judgments about each query. These analyses will test the first points of differentiation between the models: their different predictions with respect to conditional probability judgments and their specific parameterizations.

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 Implementation of Participant-Level Query-Averaged Response Models

The PT+N model defines expected probability judgments (P_e) as:

$$\begin{aligned} P_e(A) &= (1-2d)P(A)+d \\ P_e(A\wedge B) &= (1-d')P(A\wedge B)+d' \\ P_e(A\vee B) &= (1-d')P(A\vee B)+d' \\ P_e(A|B) &= \frac{(1-2d)^2P(A\wedge B)+d(1-2d)\big(P(A)+P(B)\big)+d^2}{(1-2d)P(B)+d} \end{aligned} \tag{8}$$

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

$$\begin{aligned} P_e(A) &= \frac{N}{N+2\beta}P(A)+\frac{\beta}{N+2\beta} \\ P_e(A\wedge B) &= \frac{N'}{N'+2\beta}P(A\wedge B)+\frac{\beta}{N'+2\beta} \\ P_e(A\vee B) &= \frac{N'}{N'+2\beta}P(A\vee B)+\frac{\beta}{N'+2\beta} \\ P_e(A|B) &= \frac{N}{N+2\beta}P(A|B)+\frac{\beta}{N+2\beta} \end{aligned} \tag{9}$$

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, $\vec{\theta}$, representing the probability of the elementary events ($A\wedge B$, $\neg A\wedge B$, $A\wedge\neg B$, $\neg A\wedge\neg B$).

Zhu et al. (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. (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 computed from $\vec{\theta}$ and the query, the noise parameters d and d', and the relevant equation as defined by the PT+N theory (see supplemental materials for implementation details). 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'. [3]

Recall that Zhu et al. (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.[4] I 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:

$$\begin{aligned} \mu_{ijk} &= f_{BS}(\overrightarrow{\theta_{jk}}, x_{ijk}, d_j, d'_j) \\ d_j &= \frac{1}{3}\,\text{logistic}(\delta_j) \\ d'_j &= \frac{1}{3}\,\text{logistic}\big(\delta_j+\exp(\Delta\delta_j)\big) \end{aligned} \tag{10}$$

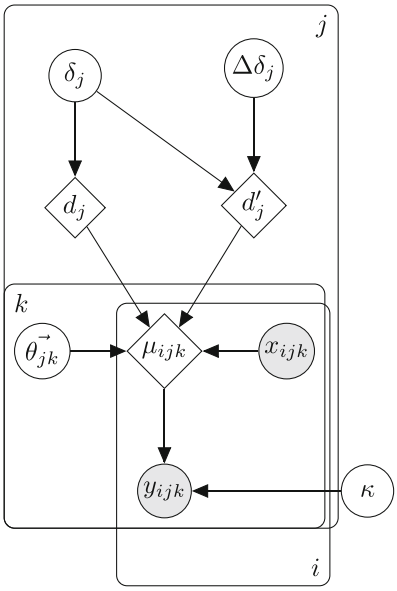
Where the function f_{BS} computes the expected probability estimate as prescribed by the Bayesian Sampler theory.

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

[3] Uninformativeness was sought in order to reduce bias in the posterior parameter estimates. It should be acknowledged that a uniform prior does not exactly correspond to what the authors of the PT+N theory would predict, as they have frequently assumed d to be a fairly small value (e.g. Costello and Watts, 2017)

[4] 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 et al. (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 $d \leq d'$. These ideas will be developed further in the trial-level analyses.

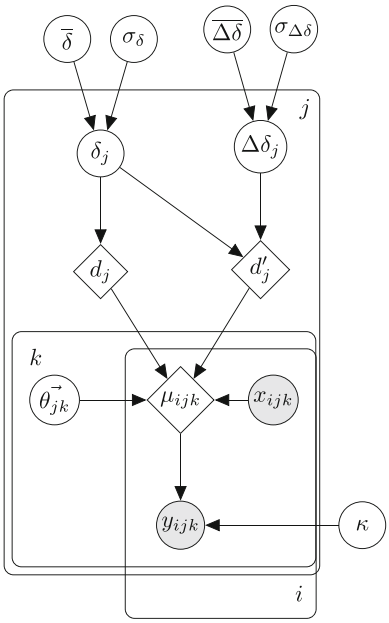


$$y_{ijk} \sim Beta(\mu_{ijk}\kappa, (1-\mu_{ijk})\kappa)$$
$$\mu_{ijk} = f_{PT+N}(\vec{\theta_{jk}}, x_{ijk}, d_j, d_j)$$
$$d_j = \frac{1}{2}\,\text{logistic}(\delta_j)$$
$$d'_j = \frac{1}{2}\,\text{logistic}(\delta_j + \exp(\Delta\delta_j))$$
$$\vec{p_{jk}} \sim Dirichlet(\vec{1})$$
$$\delta_j \sim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

Fig. 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

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. 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. 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).

Finally, I also explored fitting a hierarchical version of the Bayesian Sampler model 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 et al. (2020) acknowledge there are situations where an informative prior may be warranted (see e.g. Fennell and



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

Fig. 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

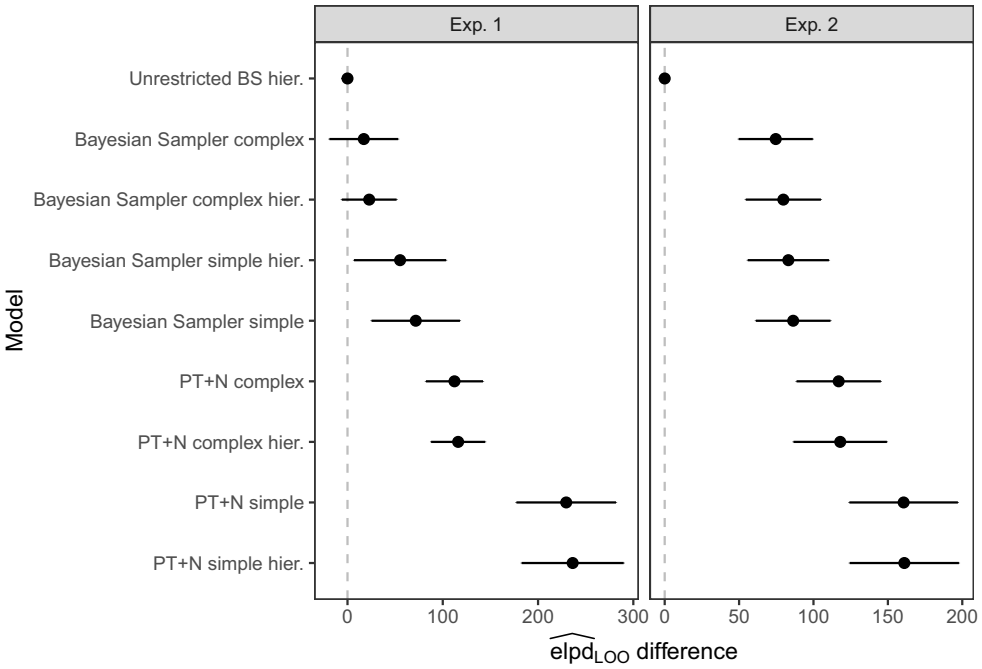


Fig. 3 Model comparison results for data from Experiments 1 and 2. Error bars indicate two standard errors of the estimates. Typically, a difference of greater than two standard errors is taken as clear evidence for the superiority of the lower-scoring model (Sivula et al., 2020)

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 $N \rightarrow 1$ and $\beta \rightarrow \infty$. 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 two-stage process of conditional probability judgment. Thus, fitting this additional unrestricted model allows for a complete comparison of the models along both of their differing dimensions.

Simulation studies verified that the complex hierarchical PT+N and Bayesian Sampler models can correctly and unbiasedly recover parameters from simulated data (see Supplemental Materials).

Model Comparison

I fit each of the models specified above to data from Zhu et al. (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

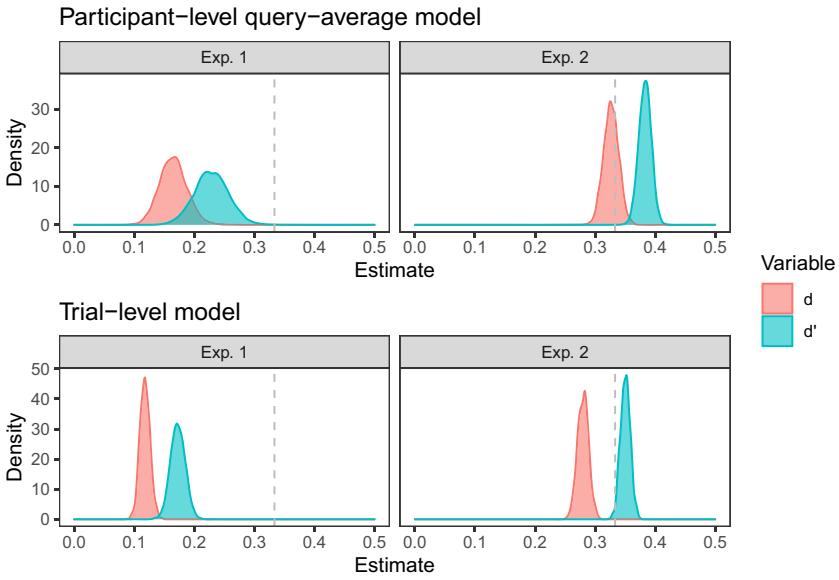


Fig. 4 Posterior density of population-level d and d' parameters estimated from the unrestricted hierarchical Bayesian Sampler model for data from Experiments 1 and 2. Dashed line indicates theoretical maximum values for Bayesian Sampler model with uninformative priors

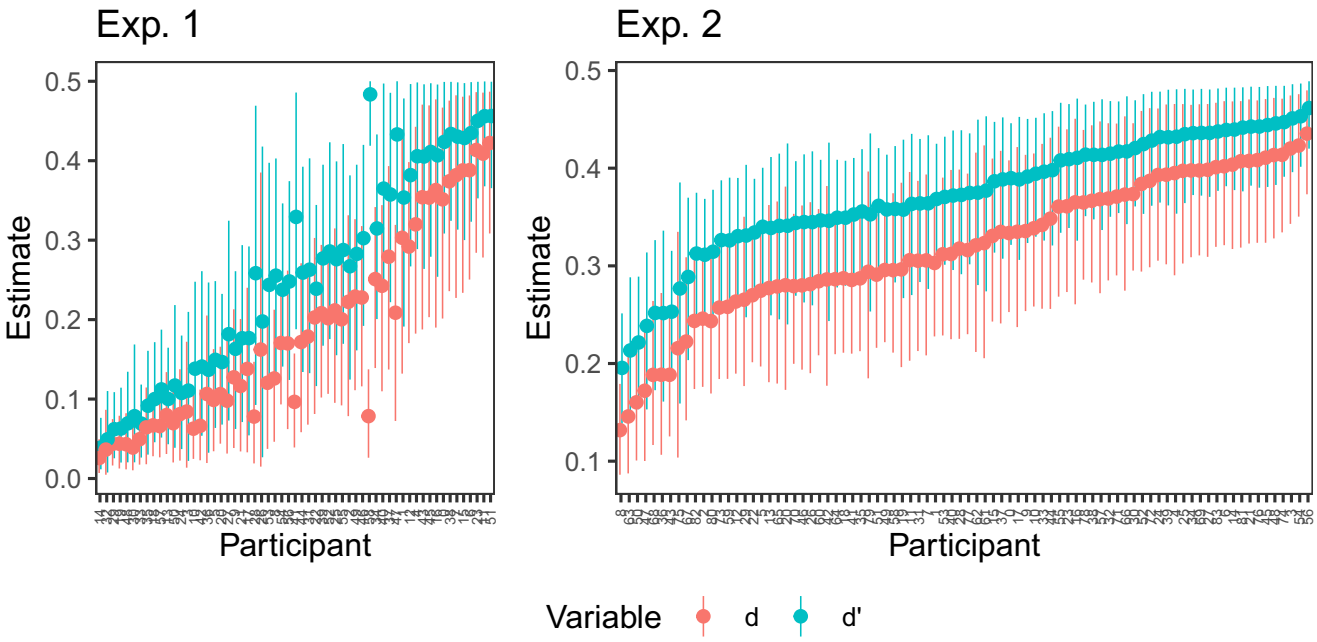


Fig. 5 Participant-level estimated d and d' values across Experiments 1 and 2. Error bars indicate 95% CIs

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 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 (greater values of $\widehat{\text{elpd}}_{\text{LOO}}$ are better). As shown in Fig. 3, the best-scoring model is an unrestricted variant of the Bayesian Sampler that allows for people to bring informative priors to the probability judgment task (i.e. allowing $\beta \in [0, \infty]$. Data from Experiment 2 more decisively reveal a single winning model: the hierarchical "unrestricted" implementation of the Bayesian Sampler model allowing for informative priors.

This unrestricted BS model differs from the PT+N model only in its treatment of conditional probability

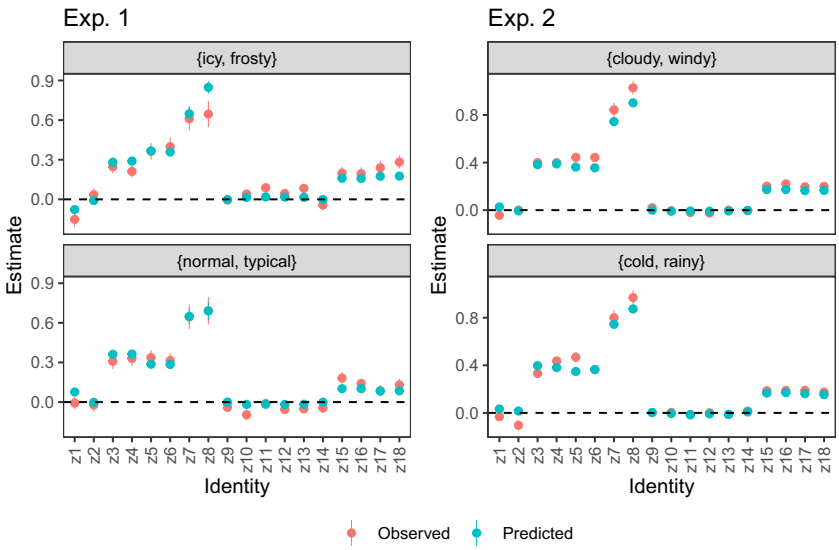


Fig. 6 Average model predicted and observed values for the 18 identities. Note that the Bayesian Sampler but not the PT+N model is capable of predicting non-zero values for identities Z10 through Z13. Error bars represent 95% CI. In Experiment 1, like participants' responses, the model's estimates are very slightly positive for {icy, frosty} and very slightly negative for {normal, typical}. This pattern replicates the qualitative pattern reported by Zhu and colleagues

judgments and so, from its superior fit, we can infer that the Bayesian Sample theory provides a better account of human conditional probability judgments.

Figure 4 (top) 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), as shown in Fig. 5. 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.

Zhu et al. (2020) demonstrated that the Bayesian Sampler model can capture a set of probabilistic identities developed by Costello and Watts (Costello and Watts, 2016, 2018) that capture some of the incoherence in people's probability judgments. Following the design of the present experiments, these identities involve combinations of probability estimates for different combinations of two events A and B that should all be equal to zero according to probability theory. Under the Bayesian Sampler and PT+N theories, however, some of these identities should be zero, but some are allowed to take on other values. Figure 6 shows the average prediction of the winning model against the average observed value for each equality. Consistent with prior findings, this model captures these identities quite closely.

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 1. Figure 7 shows the correlation between participants' responses across all trials and the best-performing model's predictions.

Results: Raw Trial-Level Response Distributions

The best-fitting model capturing participant's predictions in expectation can be seen either as a variant of the Bayesian Sampler theory allowing for informative priors or as a variant of the PT+N theory without any special treatment of conditional versus unconditional probability judgments. Thus, these prior analyses have not decisively ruled between the different process-level psychological theories behind the models. To evaluate the Bayesian Sampler and PT+N theories' competing accounts of the psychological processes behind probability judgments, a second set of analyses focused on the distribution of participant's trial-by-trial responses was conducted.

Recall, in contrast to a noise-based account, the Bayesian Sampler theory predicts both a truncation of the response distribution as well as a correlation between the degree of shrinkage in probability estimates and the variability of those estimates trial-to-trial (assuming β is relatively small). Because the Bayesian Sampler applies a Bayesian adjustment after sampling, it predicts probability judgments will always lie between d and $1 - d$, even when zero pos-

Table 1 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_{query}	$r_{\text{query-avg}}$	$\widehat{\text{elpd}}_{\text{LOO}}$	$\hat{p}_{\text{LOO}}$	r_{query}	$r_{\text{query-avg}}$
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

Models are compared based on $\widehat{extelpd}_{extLOO}$ and the correlation between their predictions for each participants' query-level average (r_{query}) as well as overall query-level averages $r_{\text{query-avg}}$

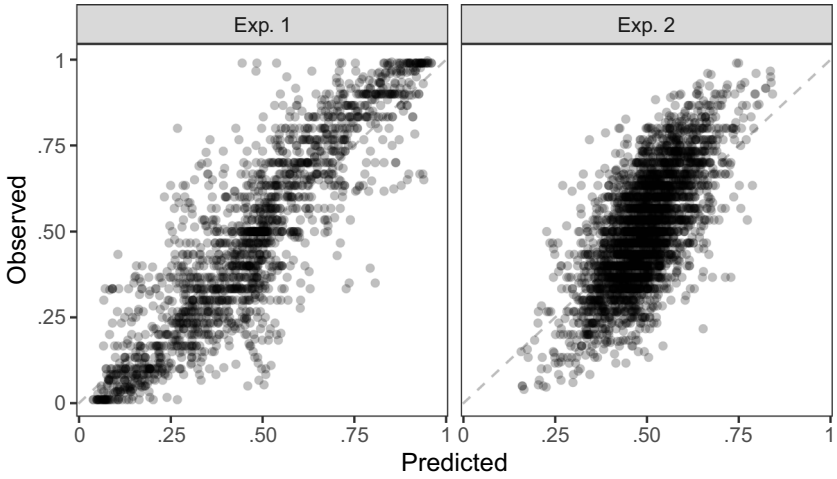


Fig. 7 Posterior predictions for best-fitting model and participants responses in Experiments 1 and 2

itive or negative mental samples are drawn (see Eq. 12 and the bridging condition, Eq. 5). First, it bears noting that the truncated response distribution implied by the Bayesian Sampler model appears at odds with the raw response data: participants' responses frequently lie outside the range implied by the best estimates of their d parameters (41% in Experiment 1 and 60% in Experiment 2).

Yet comparing the distributional predictions of the models more rigorously poses three challenges: (1) the discrete nature of the models suggests a limited set of allowable responses, assigning zero probability to all others, (2) Bayesian adjustment implies truncation of the support of the response distribution, again assigning zero probability to other response values, and (3) participants routinely round their responses, complicating both of the previous issues.

In the following set of analyses I attempt to lay out a set of reasonable assumptions that permit participants' trial-by-trial responses to be modeled and used to compare the theories' predictions. To do so, I first extend the models so as to render them fully continuous in their latent space and then marry them with a specific model of response errors. Thus the models compared in the following analyses are not identical to those originally proposed by Zhu et al. (2020) and Costello and Watts (2014). However, they do provide implementations of the theories' process-level accounts, and thus a means to test the distributional predictions of models based on Bayesian adjustment against models based on sampling noise.

Continuous Extensions of the Models

Under both models, the variability of people's responses trial-to-trial is driven by the number of mental samples drawn: more mental samples produce less-variable responses. However, if the number of samples is considered to be a truly discrete quantity, then only a limited number of discrete responses are possible. As Zhu et al. (2020) note, this is somewhat implausible on its face and their later work has abandoned this assumption (Zhu et al., 2021). At the same time, from a pragmatic perspective it is highly desirable that all latent parameters within the models be continuous rather than discrete. The models would be far more tractable to fit if, rather than including a latent Binomial variable representing the discrete number of samples drawn, we could instead model a continuous proportion of samples using, for instance, a Beta distribution.

Zhu et al. (2020) introduce the possibility of an "autocorrelated" Bayesian Sampler model under which samples are assumed to be autocorrelated (ideas which were advanced further in Zhu et al., (2021)). As autocorrelated samples provide less information than i.i.d samples, they should be weighted when computing probability estimates. The idea is that people actually draw N autocorrelated samples that approximate some smaller effective number of i.i.d samples (N_{eff}). Assuming the actual number of autocorrelated samples drawn, N, is allowed to vary somewhat noisily, then a model based on autocorrelated samples would no longer be limited to predicting a discrete set of possible responses. To approximate this as part of a wholly continuous model, I model the proportions calculated from such a hypothetical autocorrelated sampling process using a Beta distribution.

Mixture Modeling: Rounding and Contaminants

Creating continuous extensions of the models makes their estimation more tractable. However, a specific model of participants' response processes and errors is still needed to capture rounded and out-of-support responses. To address these challenges, I implemented variants of the PT+N and Bayesian Sampler models within discrete mixture models allowing for varying rounding policies as well as "contaminant" responses generated by noise processes outside the models.

First, it is clear that participants have rounded a majority of their responses. This sort of rounding can be modeled by a categorical distribution across the discrete possible rounded responses. Each rounded response category corresponds to a set of cut points, a and b, with the probability of the categorical response defined by the cumulative distribution function of the underlying latent distribution (B).

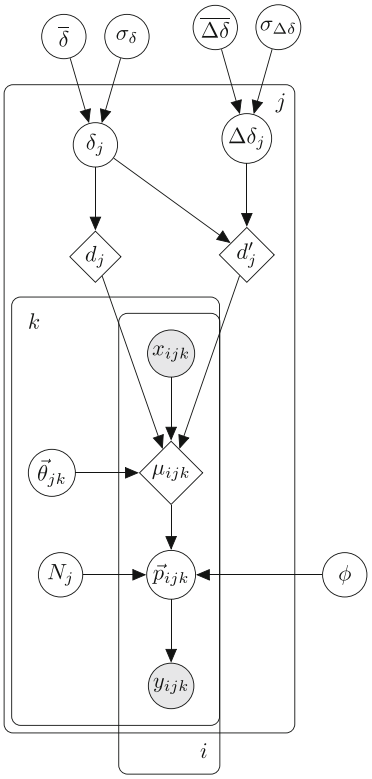
$$P([a,b)) = B(a, \mu N, (1-\mu)N) - B(b, \mu N, (1-\mu)N) \tag{11}$$

As participants were allowed to respond freely with whole numbers from 0 to 100, the exact rounding policy for each response is unknown. Nevertheless these rounding policies can be estimated via mixture modeling. For simplicity, rounding to the nearest 5% was enforced for all responses. Then, the probability of these categorical responses are computed for 21 and 11 categories (corresponding to rounding to 5% and 10%). These probabilities were combined along with a uniform probability representing "contaminants" according to mixing probabilities ϕ, distributed with a Dirichlet prior (see Appendix for further implementation details).

The Bayesian Sampler model predicts a truncated range of possible responses given β and effective N (and consequently, implied d). Modeling these different rounding processes allows for at least some out-of-bounds responses to be accounted for by rounding processes (e.g. when an allowable response of .14 is rounded to the out-of-bounds value of .10). However, some responses still cannot be accounted for by the model. Instead, these responses are treated as "contaminants" generated by a random response process. Modeling "contaminant" response processes allows the Bayesian Sampler model to be fit in the presence of true outliers. Identifying the estimated proportion of "contaminant" responses can also provide a check on the models: if a model can only be fit by assuming a large proportion of contaminant responses, this suggests it is likely not a good model of human behavior.

Trial-Level Noise-Based Model

Compared to the query-averaged model, the trial-level noise-based model adds two features: mixture components for rounding and contaminants and subject-level varying N in place of a fixed K parameter. This model's implementation and the implementation of its mixture components is depicted in Fig. 8. However, note that N is allowed to vary independently from d, allowing for independence between response shrinkage and variability.



$$y_{ijk} \sim Categorical(\overrightarrow{\theta_{ijk}})$$
$$\overrightarrow{p_{ijk}} = Z_{NB}(\mu_{ijk}, N_j, \phi)$$
$$\mu_{ijk} = f_{NB}(\overrightarrow{\theta_{jk}}, d_j, d'_j, x_{ijk})$$
$$d_j = \frac{1}{2}\,\text{logistic}(\delta_j)$$
$$d'_j = \frac{1}{2}\,\text{logistic}\Big(\delta_j + \exp(\Delta\delta_j)\Big)$$
$$\delta_j \sim Normal(0,1)$$
$$\Delta\delta_j \sim Normal(0,1)$$
$$log(\sigma_\eta) \sim Normal(-.5,.5)$$
$$log(\sigma_{\Delta\eta}) \sim Normal(-.5,-.5)$$
$$log(\sigma_\beta) \sim Normal(-1,.5)$$
$$\overrightarrow{\theta_{jk}} \sim Dirichlet_4(\overrightarrow{1})$$
$$\overrightarrow{\phi} \sim Dirichlet_3(\overrightarrow{1})$$
$$N_j \sim Cauchy(20)$$

Fig. 8 Hierarchical complex trial-level noise-based model diagram and formula specifications. Z_{NB} and f_{NB} are functions that compute the probability of each categorical response and the expected proportion of read-out mental samples given underlying mental probabilities and sample reading noise. See Appendix for further descriptions of these details

Trial-Level Bayesian Sampler Model

Approximating an autocorrelated sampling process using a Beta distribution and starting from the original Bayesian Sampler model,

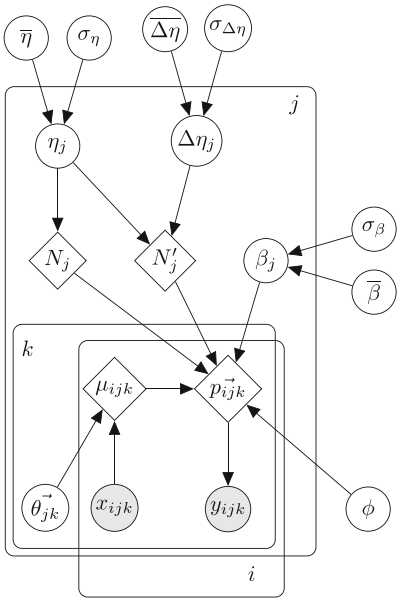
$$\hat{P}_{BS}(A) = \frac{S(A)}{N+2\beta} + \frac{\beta}{N+2\beta} \tag{12}$$

we can replace the number of successes $S(A)$ (distributed binomial) with the quantity $\rho(A)N$, where $\rho(A)$ represents the Beta-distributed sample proportions generated by the autocorrelated sampling process outlined above.

$$\hat{P}_{BS}(A) = \frac{\rho(A)N}{N+2\beta} + \frac{\beta}{N+2\beta} \tag{13}$$

Then it is plain that $\hat{P}_{BS}(A)$ is a transformation of $\rho(A)$, and therefore a transformed Beta distribution (see Appendix for derivation). Figure 9 diagrams the entire Bayesian Sampler model, now parameterized in terms of β, N, and N'.

This model assumes that the number of samples drawn on each trial is fixed as N or N' accordingly (modulo the uncertainty about these parameters). However, it also seems reasonable to imagine that the number of samples drawn in fact varies from trial-to-trial. For the Bayesian Sampler model, this could substantially impact the model's fit. As the number of samples drawn affects the truncation of the response distribution, this may allow the model to capture some responses that would otherwise be treated as contaminants. To capture this, the model can be given one additional extension to allow the number of samples to vary, by adding a new parameter N_{trial}. This parameter multiplies the number of effective samples as a fraction of each individual participants' average number of samples drawn, e.g. so that a participant might sometimes draw 1.5× or 2× the number of effective samples they typically draw. The appropriate amount of variation in N_{trial} is constrained to be fairly small, but is estimated hierarchically: I assume $log(N_{trial}) \sim N(0, \sigma_{N_{trial}})$ and $log(\sigma_{N_{trial}}) \sim (-1, .3)$.



$$y_{ij} \sim Categorical(\theta)$$
$$\overrightarrow{\theta_{ijk}} = Z_{BS}(\mu_{ijk}, N_j, N'_j, \beta_j, \phi)$$
$$\mu_{ijk} = f_0(\overrightarrow{\theta_{jk}}, x_{ijk})$$
$$N'_j = 1 + exp(\eta_j)$$
$$N_j = 1 + exp(\eta_j) + \exp(\Delta\eta_j)$$
$$log(\beta_j) \sim Normal(\overline{\beta}, \sigma_\beta)$$
$$\eta_j \sim Normal(\overline{\eta}, \sigma_\eta)$$
$$\Delta\eta_j \sim Normal(\overline{\Delta\eta}, \sigma_{\Delta\eta})$$
$$\overline{\eta'} \sim Normal(1, 1)$$
$$\overline{\Delta\eta} \sim Normal(0, .1)$$
$$\overline{\beta} \sim Normal(-.5, .4)$$
$$log(\sigma_\eta) \sim Normal(-.5, .5)$$
$$log(\sigma_{\Delta\eta}) \sim Normal(-.5, -.5)$$
$$log(\sigma_\beta) \sim Normal(-1, .5)$$
$$\overrightarrow{p_i} \sim Dirichlet(\overrightarrow{1})$$
$$\overrightarrow{\phi} \sim Dirichlet(\overrightarrow{1})$$

Fig. 9 Hierarchical complex trial-level Bayesian Sampler model diagram and formula specifications. Z_{BS} and f_0 are functions that compute the probability of each categorical response and the expected proportion of mental samples given underlying mental probabilities before Bayesian adjustment. See Appendix for further descriptions of these details

Model Comparison: Raw Trial-Level Response Models

Prior to fitting the models, response data were rounded to the nearest 5%. Nearly all responses (Exp. 1: 93% and Exp. 2: 89%) were already divisible by five, and this was necessary to speed model fitting. Even still, estimating model posteriors for the trial-level mixture models using MCMC proved intractable. Instead, model posteriors were estimated using Stochastic Variational Inference (SVI) with Numpyro (Phan et al., 2019) using a multivariate Normal guide (e.g. Kucukelbir et al., 2015). Estimating each model posterior took between approximately 20 to 90 minutes on an Nvidia V100 GPU. Simulation studies verified that this estimation approach could reliably recover parameters from simulated data (see Supplemental Materials).

Table 2 shows the scores of each model fit to the trial-level data in Experiments 1 and 2. From the quantitative model comparison it is clear that the noise-based model is superior, with substantially lower $\widehat{\text{elpd}}_{\text{LOO}}$ than both of the competing Bayesian Sampler implementations in both experiments.

The models' performance can be better understood by examining the two main features about which the models' predictions depart: truncation of the response distribution and shrinkage-dependent variance of responses.

The non-varying Bayesian Sampler model fits quite poorly and is estimated to have a large proportion of "contaminant" responses (43% in Experiment 1, 26% in Experiment 2). This is likely due to a lack of the predicted truncation of the response distribution. As with the trial-average models, the estimates of implied d are quite high, which would predict substantial truncation. As mentioned earlier, many of participants' responses fall outside the range implied by their most likely implied d values.

Allowing the effective number of samples drawn to vary trial-by-trial improves the fit of the Bayesian Sampler model substantially and somewhat decreases the estimated proportion of contaminant responses (30% in Experiment 1, 28% in Experiment 2). Nevertheless, these results still indicate inferior fit compared with the noise-based model, which has substantially lower $\widehat{\text{elpd}}_{\text{LOO}}$ and attributes far fewer responses to the contaminant process (11% in Experiment 1, 4% in Experiment 2).

Second, we can examine the shrinkage-variance relationship. In the noise-based model, N and d were allowed to vary freely. But if these quantities are actually correlated as predicted by a Bayesian adjustment model, then we should expect to nevertheless see a correlation between the subject-level estimates in these parameters. Figure 10 shows scatterplots relating these estimates. Although there is a slight negative correlation in Experiment 1, this is driven by only a handful of participants with extreme values. In Experiment 2 the correlation appears to if anything be positive rather than negative. These findings again run counter to the predictions of the Bayesian Sampler theory.

Comparing the d values inferred from the trial-level model we see they are similar to but generally smaller than those estimated in the trial-averaged model. It is unclear exactly why this is, though it could owe at least in part to the mixture component preventing "contaminant" responses from affecting the estimates of these parameters.

Discussion

Fit to the average of participants' responses over blocks, there is a single clear winner among the competing models: a model without any special treatment of conditional probability (ala the Bayesian Sampler model) and allowing for an implied d parameter $\in [0, .5]$. This model could be interpreted either as a variant of the Bayesian Sampler without restriction on its β parameter, or as a variant of the PT+N model that removes its account of conditional probability judgments.

In either case, these findings make clear 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

Table 2 Bayesian model comparison results for trial-level models with best scoring model in bold face

Experiment	Model	$\widehat{\text{elpd}}_{\text{LOO}}$	$\hat{p}_{\text{LOO}}$	$\widehat{\Delta\text{elpd}}_{\text{LOO}}$	$\text{SE}_{\Delta\text{LOO}}$
Exp. 1	**Noise-based**	**–12256.0**	**437.9**	**0.0**	**0.0**
	BS: Varying N	–14079.7	1617.8	1823.7	74.5
	Bayesian Sampler	–14866.7	1384.5	2610.7	83.8
Exp. 2	**Noise-based**	**–24053.4**	**574.0**	**0.0**	**0.0**
	Bayesian Sampler	–25376.0	1177.3	1322.6	56.0
	BS: Varying N	–26027.5	2154.6	1974.1	53.9

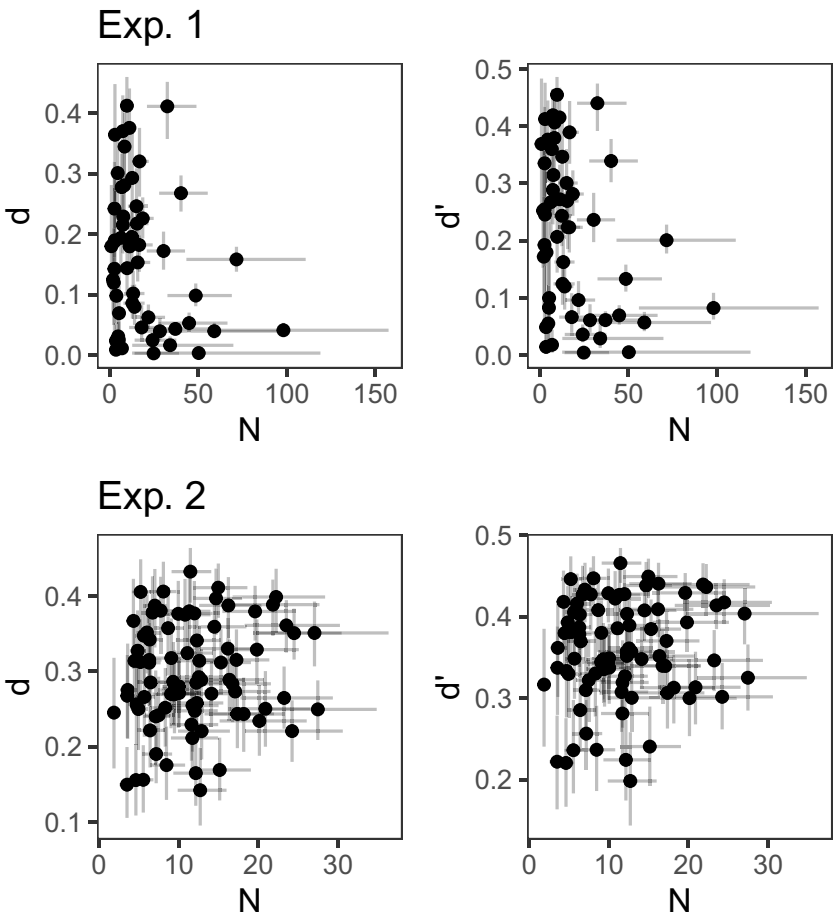


Fig. 10 Scatterplots showing the relationships between subject-level d and N estimates from the trial-level noise-based model for Experiments 1 and 2. Figure for Experiment 1 excludes some outlier paricipants who gave repetitive responses that resulted in abnormally high N-values. Error bars indicate 95% CI

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).

At the psychological process-level, the Bayesian adjustment process hypothesized by the Bayesian Sampler model makes two clear predictions about the distribution of participants' responses. First, it implies a truncated range of possible responses. Second, assuming that β is constrained to be a relatively small value, then under the Bayesian Sampler model, N influences both the degree to which responses are shrunk toward .50 and the variability of those responses trial-to-trial. Thus, participant's inferred N values should be somehow correlated with the noise in their trial-level responses. In contrast, under the noise-based account of the PT+N model, there is no truncation of responses and no predicted correlation between shrinkage and response variability.

The distributions of participants' responses are more consistent with the PT+N theory's account of sampling noise than the Bayesian adjustment implied by the Bayesian Sampler theory as neither of the Bayesian Sampler's predictions appear to be borne out by the data. First, participants' responses frequently fall outside the truncated range implied by the parameters estimated under the Bayesian Sampler model. Fit to the raw data, this requires treating an unreasonably large proportion of responses as "contaminants". Second, the degree of shrinkage in participants' responses and the variability in those responses are not correlated in the ways predicted by the Bayesian Sampler theory.

In the end, I find the best overall account of participants' probability judgments is a modification of the PT+N theory without its two-stage process of conditional probability judgments. Like the PT+N theory, this model accounts for distortions in probability judgments via a process of noisy sampling. But like the Bayesian Sampler theory, under this theory, conditional probability judgments are produced by a process of conditioning in the mental model of the events, rather than as part of the mental sampling process itself.

Outside of probability judgments, Bayesian conditioning is a key aspect of the sorts of mental models imagined by Bayesian cognitive scientists, where cognitive models are

conditioned on information as part of learning, prediction, and inference. Within this framework, it seems only natural to imagine that a similar conditioning process would subserve probability judgments, rather than conditional probability judgments being made by a distinct two-stage sampling process.

To illustrate the distinction, consider being asked to judge the conditional probability that it will rain tomorrow in London given it rained today in London. Now, compare that with first being told that it rained today in London, and then being asked to make the simple probability judgment that it will rain tomorrow in London. The original PT+N theory would draw a distinction between the two tasks: The first task would invoke the two-stage sampling process, whereas presumably the second would involve some change in the mental model to reflect learning about the day's weather followed by only a single stage of sampling. In contrast, the present findings suggest these tasks would invoke the same set of mental processes—conditioning of the mental model followed by the drawing of samples from that model.

Remaining Questions and Limitations

Despite the model's quantitative success, some more qualitative questions remain. First, the plausibility of the parameter values inferred from the model bears consideration. Many participants' estimated d and d' parameters were quite high—potentially against the spirit of the original PT+N model. This model bounds d at .50 in principle, but a sample-reading process with such high error-rates may or may not be plausible. In prior work, simulations have often assumed values of d around .10 (Costello & Watts, 2017; e.g. Howe & Costello, 2020). Further research examining what factors might affect the mental sampling and reading processes (e.g. task complexity, distractions, prior experience) might help to shed light on the most plausible range of d values in different contexts.

High estimates of d parameters might also call into question arguments for the rational utility of such a process. Zhu and colleagues argue that the regularizing effect of Bayesian adjustment should be seen as adaptive. They also consider that "noise" might give an algorithmic-level solution to the computation-level goals defined by the Bayesian sampler (Zhu et al., 2020). Even high implied d values might be consistent with rational inference in cases where the number of effective mental samples is very low. For instance, a Beta (2,2) prior is only modestly informative, but could produce $d = .40$ if $N = 1$. However, in cases where more samples are drawn, high d values would correspond to potentially inflexible and suboptimal priors. From the results, it is clear there are some individuals with both relatively high d and N values, which may press somewhat against the rational justification for the desirability of sampling noise. For instance, some participants are inferred to have $N \approx 20$ and $d \approx .30$, implying $\beta \approx 15$.

Finally as noted, the comparisons of these models using trial-level data rests on a number of elaborating assumptions to support fitting of the models. It should be recognized that different assumptions may have produced different results, and other error processes remain possible. Although other indirect analysis approaches might be designed to avoid these concerns (e.g. Sundh et al., 2021), ultimately it seems crucial that cognitive models at some point be fit to the actual human behaviors of interest. An important direction for future elaborations of sampling-based theories are more rigorous theories of realistic mental sampling processes, including details of their initialization, autocorrelation, and amortization (Gershman & Goodman, 2016).

Conclusions

Probability judgments have proven a fruitful testing ground for sampling-based theories of cognition. But, the implications of sampling-based models like the Bayesian Sampler and PT+N theory go well-beyond the probability judgment task itself: these models have the potential to extend the success of Bayesian theories of cognition to develop a probabilistic science of everyday beliefs. Under such an account, beliefs are not explicitly represented, stored, or even computed as probabilities, but rather they are emergent properties of mental models generating probabilistic samples (Chater et al., 2020; Sanborn & Chater, 2016).

Nevertheless we might still use the logic of Bayesian models to understand the operation of these beliefs and how they respond to evidence. Indeed, by representing the "true" subjective probabilities as a latent variable in the models used here, Bayesian data analysis allows those underlying credences to be inferred. 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, 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, 2022, Powell et al., 2018).

Appendix

For the trial-level response models participants' rounded responses are modeled as discrete responses with a categorical (multinomial) distribution. For $i \in \{0, 1, ..., m\}$ where $m = 20$ possible responses, define a set of cut points $a_i = \frac{i}{m} - \frac{1}{2m}$ and $b_i = \frac{i}{m} + \frac{1}{2m}$. Using $x|_{[0,1]}$ to denote that x is restricted to the domain [0, 1], the probability of each response given μ and N is:

$$\begin{aligned} p_{i,5} &= P([a_{i,5}, b_{i,5})) \\ &= B(a_{i,5}|_{[0,1]}, \mu N, (1-\mu)N) \\ &\quad - B(b_{i,5}|_{[0,1]}, \mu N, (1-\mu)N) \end{aligned}$$

where B is the appropriate cumulative distribution function. To capture rounding to 10 we define $a_{i,10} = \frac{2i}{m} - \frac{1}{m}$ and $b_i = \frac{2i}{m} + \frac{1}{m}$, so that the probability of each response is:

$$p_{i,10} = \begin{cases} P([a_{i,10}, b_{i,10})) & \text{i is even} \\ 0 & \text{i is odd} \end{cases} \tag{14}$$

Next defined a vector of mixture probabilities $\overrightarrow{\phi}$, with the zeroth index indicating a "contaminant" process. Combining these response processes, we can define the marginal probability of each response as:

$$p_i = \frac{1}{21}\phi_0 + p_{i,5}\phi_1 + p_{i,10}\phi_2$$

And then responses themselves are distributed Categorical:

$$y_i \sim Categorical(\overrightarrow{p})$$

For the noise-based model, B is the incomplete Beta function, the CDF of the Beta distribution. The computations for the Bayesian Sampler model response probabilities are identical save that instead of $B(x, \alpha, \beta)$ we have $B(f_{BS}^{-1}(x), \alpha, \beta)$ when computing the probability of each response p_i, and we use N and N' where appropriate. To see this, let X be the Beta distribution success proportions from the mental sampling operations, $\rho(A)$, and let Y be the distribution of resulting probabilities from the Bayesian sampler model. Then $Y = g(X)$ where g is the function defined in equation 13 from the manuscript.

$$\hat{P}_{BS}(A) = \frac{\rho(A)N}{N + 2\beta} + \frac{\beta}{N + 2\beta} \tag{15}$$

Letting F_X and F_Y be the CDF of X and Y respectively, we have that:

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) \end{aligned}$$

Putting this all together, define Z_{NB} as the function which calculates the probability of each categorical response under the noise-based model given the inputs of μ_{ijk}, d_j, d'_j and ϕ. Here, $\mu_{ijk} = f_{\text{NB}}(\overrightarrow{\theta_{jk}}, d_j, d'_j, x_{ijk})$ computes the expected probability according to the PT+N theory except that it treats conditional probability judgments like simple probability judgments.

Finally, define Z_{BS} as the function which calculate the probability of each categorical response under the Bayesian Sampler model. Note that, here, $\mu_{ijk} = f_0(\overrightarrow{\theta_{jk}}, x_{ijk})$ where the value of μ depends only on the underlying probabilities and query asked on a specific trial.

Author Contribution Derek Powell is the sole author of this manuscript.

Availability of Data and Materials This paper presents secondary analyses of data. The datasets generated and/or analysed during the current study are available at https://osf.io/mgcxj/files/.

Code Availability All analysis code is available at https://github.com/derekpowell/bayesian-sampler and at https://osf.io/bpkjf/.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflict of Interest The author declares no competing interests.

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.

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.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.) *Formal representation of human judgment, pp 17–52, New York, Wiley.*

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. https://doi.org/10.1037/0033-295X.101.3.519.

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...., & Goodman, N.D. (2016). What does the crowd believe? A hierarchical approach to estimating subjective beliefs from empirical data, vol. 6.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D.B. (2014). Bayesian data analysis (Third edn.). Boca Raton: 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.

Gershman, S. J., & Goodman, N. D. (2016). Amortized inference in probabilistic reasoning. vol. 7.

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.

Howe, R., & Costello, F. (2020). Random variation and systematic biases in probability estimation. *Cognitive Psychology*, *123*, 101306. https://doi.org/10.1016/j.cogpsych.2020.101306.

Jaynes, E. T. (2003). Probability theory: The logic of science (G. L. Bretthorst, Ed.) Cambridge, United Kingdom: Cambridge University Press.

Jern, A., Chang, K. 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 edn.) New York: 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.

Kucukelbir, A., Ranganath, R., Gelman, A., & Blei, D. (2015). Automatic variational inference in stan. In *Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2015/hash/352fe25daf686bdb4edca223c921acea-Abstract.html*.

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

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. Statistical Science, vol 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. (2022). A descriptive Bayesian account of optimism in belief revision. In C. Jennifer, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.) *Proceedings of the 42nd annual conference of the cognitive science society*.

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

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. https://doi.org/10.1016/j.tics.2016.10.003.

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.

Sundh, J., Zhu, J., Chater, N., & Sanborn, A. (2021). The mean-variance signature of Bayesian probability judgment. PsyArXiv. https://doi.org/10.31234/osf.io/yuhaz.

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

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, vol. 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.

Zhu, J.-Q., Sundh, J., Spicer, J., Chater, N., & Sanborn, A. (2021). The autocorrelated Bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. PsyArXiv. https://doi.org/10.31234/osf.io/3qxf7.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.