Comparing probabilistic accounts of probability judgments

Derek Powell[1]

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

Author Note

This manuscript is pre-print that has not yet been peer reviewed.

Abstract

Bayesian theories of cognitive science hold that cognition is fundamentally probabilistic. However, one mark sometimes held against these theories is that 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, Sanborn, & Chater, 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. Compared 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 that allows for informative priors.

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

Word count: X

Comparing probabilistic accounts of probability judgments

Bayesian theories of cognition offer a unified formal framework for cognitive science (Tenenbaum, Kemp, Griffiths, & Goodman, 2011) that has had remarkable explanatory successes across domains, including in perception (e.g. Kersten, Mamassian, & Yuille, 2004), memory (Anderson, 1991), language (Xu & Tenenbaum, 2007), and reasoning (Lu, Chen, & Holyoak, 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 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."

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 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 variety of other findings illustrating the incoherence of human probability judgments (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, Sanborn, & Chater, 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, Schulz, & Gershman, 2017).

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 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:

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

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 accounts 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 parameterized by β. 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. Since for any Beta distribution $x \sim Beta(a, b)$, $E[x] = \frac{a}{a+b}$, the probability estimate is thus 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, the effect of a Bayesian prior is nearly identical to the effect of noise in the PT+N model. 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, how they treat 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:

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

The PT+N's account of conditional probability judgment is distinctly non-Bayesian, separating it 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 fallacies—key findings in the probability judgment literature (Costello & Watts, 2017; Zhu, Sanborn, & Chater, 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, Sanborn, & Chater, 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 fallacies are an extremely robust empirical finding (Mellers, Hertwig, & Kahneman, 2001; Sides, Osherson, Bonini, & Viale, 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) elect to split the baby by averaging the simple and complex model scores together. This approach somewhat favors the Bayesian Sampler theory overall, though they are cautious about drawing any strong conclusions in favor of their theory.

Accounting for Model Complexity. When comparing and selecting models based on their predictive accuracy, it is necessary to correct for the potential for these models to "overfit" the 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 models' complexity.

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 [0, 1/3] (approaching 1/3 as N approaches ∞ and β approaches 1), whereas the PT+N model permits noise values between $[0, 1/2]$.

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.

Finally, 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.

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 on information criteria, such as $WAIC$ and LOO_{ic}. Both of these indices approximate leave-one-out cross validation, thereby offering a measure of the ability of a model to capture new unseen data. Importantly, these indices also allow for complexity penalties that directly capture differences in the flexibility of the models, instead of being estimated based solely on the number of distinct parameters (Gelman, Hwang, & Vehtari, 2014; Vehtari, Gelman, & Gabry, 2017). In addition, the Bayesian framework also 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 a potential reduction 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 London 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 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 both the Bayesian Sampler and PT+N models as Bayesian models 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 probability judgments as:

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

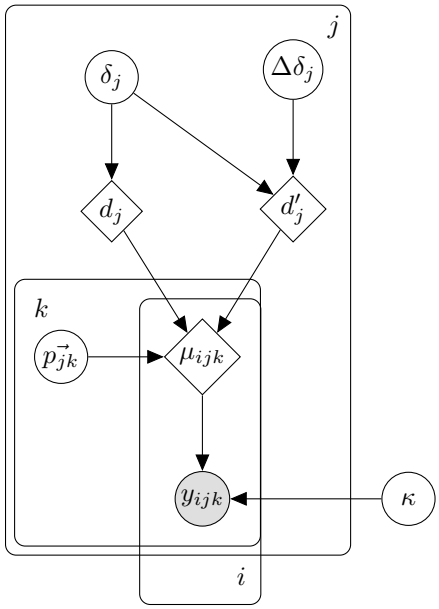
In contrast, the Bayesian Sampler model defines probability judgments as:

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

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. Figure 1 displays the translation of the PT+N model into the Bayesian framework, along with a plate diagram representing the dependencies among parameters.



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

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.

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).

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. 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.

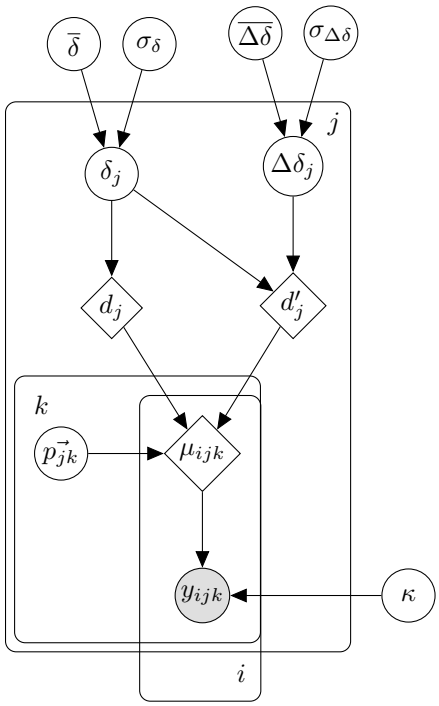
The Bayesian Sampler model is therefore identical to the PT+N model save for the changes to μ_{ijk}, d, and d' shown below:

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

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 should prevent over-penalizing complex versions of the models and potentially improve their overall predictive fit. 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, Roberts, & Sköld, 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.



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

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.

Finally, I also explored fitting versions 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, Sanborn, & Chater, 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 and β both approach ∞. 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 computed LOO_{ic} scores for each combination. Compared with BIC, LOO_{ic} offers more sophisticated methods for estimating model complexity, and are generally considered 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, Simpson, Yao, & Gelman, 2019). Model posteriors were estimated using the Numpyro (Phan, Pradhan, & Jankowiak, 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 LOO_{ic} 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 is no clear winner. The beset-scoring model is an unrestricted variant of the Bayesian

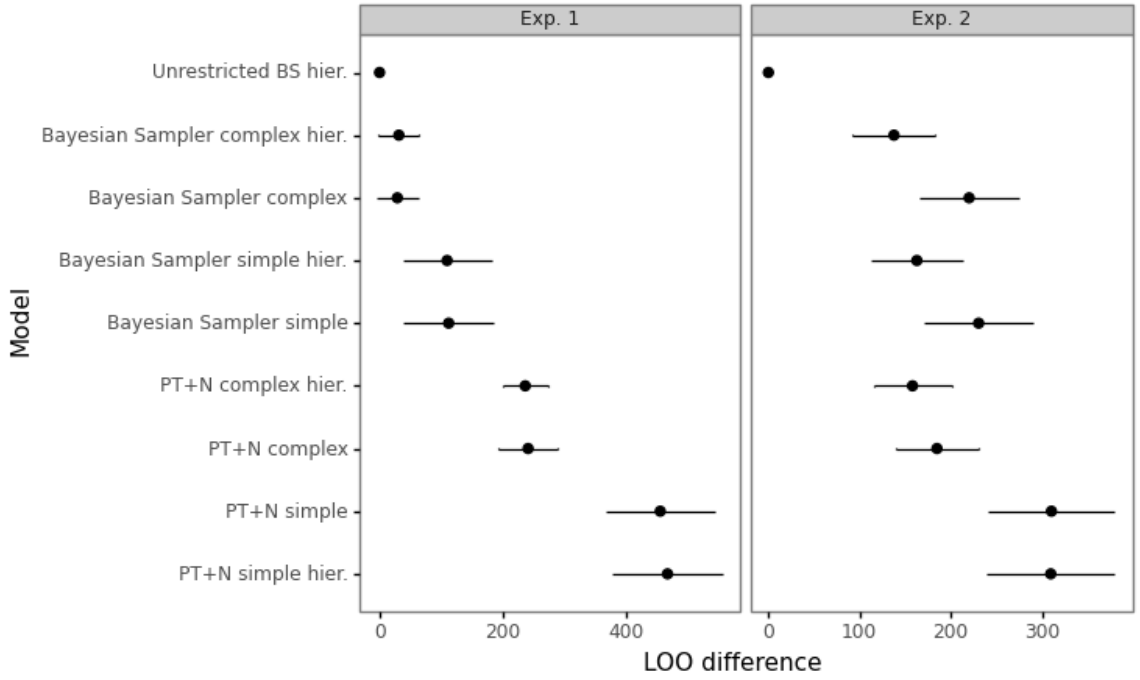


Figure 3. Model comparison results for data from Experiments 1 and 2. Error bars indicate two standard errors of the estimates. Typically, a difference of greater than two standard errors is taken as clear evidence for the superiority of the lower-scoring model [**?**].

Sampler allows for people to bring informative priors to the probability judgment task (again, lower scores for LOO_{ic} are better). That is, a Bayesian Sampler model without any restriction on β, allowing it to range from [0, ∞]. However, the complex and complex hierarchical implementations of the Bayesian Sampler assuming uninformative priors (i.e., restricting $\beta \in [0, 1]$) have LOO_{ic} scores within two standard errors of the difference, indicating that these models are also plausible.

Data from Experiment 2 more decisively reveal a single winning model: the hierarchical "unrestricted" implementation of the Bayesian Sampler model allowing for informative priors. Recall, this model may also be seen as a version of the PT+N theory that removes the original theory's constructive account of conditional probability judgments. Comparing predictions for conditional probability judgments 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). These findings suggest that conditioning is better seen as part of the mental model than the probability judgment process.

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 and d' are greater than $\frac{1}{3}$, 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 (13 of 59 for d, 26 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.

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'. Evidently this structural component isn't a complexifying epicycle, instead it is better seen as a prediction that constrains flexibility. However, it is one that ultimately 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, but it illustrates how 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

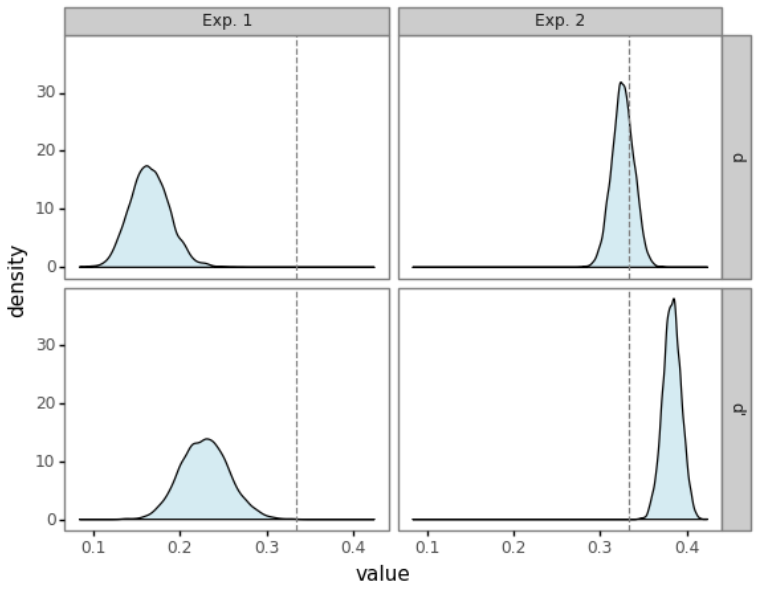


Figure 4. Posterior density of population-level d and d' parameters estimated from the unrestricted hierarchical Bayesian Sampler model for data from Experiments 1 and 2. Dashed line indicates theoretical maximum values for Bayesian Sampler model with uninformative priors.

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 for this data under this model are both very near 1/3, 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. 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, Simpson, Yao, &

Gelman, 2019). Fortunately, conclusions from the comparisons here do not rest solely on differences between the models' complexity.

Table 2

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

	Experiment 1				Experiment 2			
Model	LOO_{ic}	Penalty	r_{resp}	r_{trial}	LOO_{ic}	Penalty	r_{resp}	r_{trial}
Unrestricted BS hier.	**-2163.2**	**296.4**	**0.884**	**0.964**	**-3953.3**	**368.4**	**0.688**	**0.878**
Bayesian Sampler complex	-2134.4	285.1	0.883	0.961	-3733.2	453.6	0.679	0.848
Bayesian Sampler complex hier.	-2131.9	292.0	0.883	0.960	-3815.6	399.5	0.675	0.852
Bayesian Sampler simple hier.	-2053.9	271.8	0.874	0.953	-3790.5	383.2	0.667	0.839
Bayesian Sampler simple	-2051.3	273.7	0.874	0.952	-3722.8	418.9	0.667	0.831
PT+N complex hier.	-1927.4	298.3	0.867	0.946	-3795.3	356.2	0.658	0.835
PT+N complex	-1922.8	264.5	0.864	0.941	-3768.5	398.2	0.667	0.840
PT+N simple	-1708.8	254.5	0.838	0.918	-3643.0	319.5	0.619	0.777
PT+N simple hier.	-1697.1	265.4	0.839	0.919	-3643.8	305.3	0.617	0.772
Relative Freq.	-1266.2	305.7	0.820	0.875	-1287.7	424.6	0.516	0.639

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

Discussion

By a fair margin, the model best accounting for the experimental data from Zhu and colleagues (2020) was 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 experimental task. In keeping with the larger theoretical framework of Bayesian cognitive science, this 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, Sanborn, & Chater, 2020).

Findings from Zhu and colleagues' experiments do cast 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 an appealing rational basis. Nevertheless, the data suggest that people may bring informative priors to the probability judgment task, indicated by estimated d and d' parameters outside the [1, 1/3] range implied by "ignorance" priors. It could be that people bring domain-specific priors to judgment tasks, meaning that the most appropriate priors might be dictated by the context in which they make their judgments. This could potentially account for the differences in d and d' observed across Experiments 1 and 2.

In addition, the implied d and d' noise parameters also varied across individuals. Of note, even in Experiment 2, some participant's implied d and d' parameters were consistent with the class of ignorance priors (see supplemental materials). Further research should explore this heterogeneity within and across individuals. Unfortunately, pinning down specific components of the Bayesian Sampler model is a challenge, as the β and N parameters are not uniquely identifiable from judgment data of the sort examined here.

One thing this model comparison has not decided, and likely *cannot* decide, is whether the distortions of probability judgments are products of mental noise or of further reasoning processes. Given the models' tight connections via the bridging condition (Zhu, Sanborn, & Chater, 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, Chang, & Kemp, 2014). 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, Chang, & Kemp, 2014; Powell, Weisman, & Markman, 2018).

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

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, 6.

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

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, 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 edition). 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

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]*. Retrieved from http://arxiv.org/abs/1912.11554

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

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

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

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