Word count: 10511

A unified explanation of variability and bias in human probability judgments: How computational noise explains the mean-variance signature

Joakim Sundh[1], Jian-Qiao Zhu[2], Nick Chater[3], & Adam Sanborn[2]

[1]Department of Psychology, Uppsala University

[2]Department of Psychology, University of Warwick

[3]Warwick Business School, University of Warwick

Author Note

A preliminary version of our work was presented at the 42nd Annual Virtual Meeting of the Cognitive Science Society Conference and at the 2021 meeting of the Society for Mathematical Psychology, and a preprint of this work posted on PsyArXiv (https://psyarxiv.com/yuhaz), but otherwise none of the material included in this paper has been published or is under consideration elsewhere. Informed consent was obtained from all human subjects in the experiments described in this manuscript. None of the work in this manuscript was preregistered. We have made the new data reported in this paper available here: https://osf.io/9kea6/. Zhu, Sundh, and Sanborn were supported by a European Research Council consolidator grant [817492-SAMPLING].

Abstract

Human probability judgments are both variable and subject to systematic biases. Most probability judgment models treat variability and bias separately: a deterministic model explains the origin of bias, to which a noise process is added to generate variability. But these accounts do not explain the characteristic inverse U-shaped signature linking mean and variance in probability judgments. By contrast, models based on sampling generate the mean and variance of judgments in a unified way: the variability in the response is an inevitable consequence of basing probability judgments on a small sample of remembered or simulated instances of events. We consider two recent sampling models, in which biases are explained either by the sample accumulation being further corrupted by retrieval noise (the Probability Theory + Noise account), or as a Bayesian adjustment to the uncertainty implicit in small samples (the Bayesian sampler). While the mean predictions of these accounts closely mimic one another, they differ regarding the predicted relationship between mean and variance. We show that these models can be distinguished by a novel linear regression method that analyses this crucial mean-variance signature. First, the efficacy of the method is established using model recovery, demonstrating that it more accurately recovers parameters than complex approaches. Second, the method is applied to the mean and variance of both existing and new probability judgment data, confirming that judgments are based on a small number of samples that are adjusted by a prior, as predicted by the Bayesian sampler.

Keywords: sampling, probability, biases, Bayes, noise

Public Significance Statement

Human probability judgments play a crucial role in everyday reasoning and decision-making. But it can be difficult to distinguish between different theoretical models of the mental processes determining such judgements. This study introduces a new method which uses the relationships between the mean and the variance of probability judgments to discriminate between theoretical models. Applying the method provides new evidence for a theory based on mental sampling coupled with a Bayesian adjustment of the sampled proportions, as well as a simple and accurate way to estimate model parameters for individuals. This sheds new light on many of the reoccurring biases in human probability judgment.

Introduction

During the last few decades, sampling-based models have emerged as one of the most promising accounts of human probability judgment. This perspective is based on the assumption that, when judging the probability of an event, the brain samples a number of instances from some internal representation, such as drawing them from long-term memory or performing mental simulation, and then bases the judgment on the frequencies in this sample. For example, when looking out the window and judging the probability of rain, we might sample a number of similar days and base our judgments on how many of those days it actually did rain.

Sampling-based models have been very successful at modeling a range of human behaviors; perhaps most critically, they naturally account for the stochasticity of human judgment and decision making (Bonawitz et al., 2014; Griffiths et al., 2012; Juslin et al., 2007; Sanborn, & Chater, 2016), while almost all other extant models of human probability judgment do not specifically explain this variability. Instead, other models typically add a generic error term that implicitly encompasses any type of response or measurement noise that might perturb the cognitive process, while the process itself is otherwise described as deterministic. Of course, any type of noise could potentially account for stochasticity in human probability judgment, but as we will demonstrate, different assumptions regarding the source of the noise will imply different identifiable patterns. In this perspective, the variability of human behavior is not merely a 'nuisance parameter' to be grudgingly tolerated but a source of valuable data in itself.

In this paper, we will demonstrate that, firstly, sampling-based models are associated with a particular signature pattern in the relationship between the mean and the variance of human probability judgments that cannot be accounted for by a generic additive error term and, secondly, that particular empirical characteristics of this pattern is consistent with a

Bayesian sampler account of judgment biases. As has been observed in other research (Howe & Costello, 2020; Kvam & Busemeyer, 2020; Ren et al., 2021), human probability judgments are associated with an inverse U-shaped relationship between the mean probability and the variance of judgments, where judgments near the edges of the probability scale (i.e., 0 or 1) are associated with less variability than judgments near the middle of the probability scale (i.e., .5). This is consistent with a binomial response distribution with variance $np(1 – p)$, such as is generally assumed in sampling-based models. In deterministic models however, error is generally modeled by the generic normally-distributed error term $\varepsilon \sim N(0, \sigma)$. In this case, variance is defined solely by σ, which is typically assumed to be independent from p, and therefore (aside from at the boundaries) will predict variance to be constant over the probability scale. This is notable because in a majority of extant models, such as most heuristics, noise is conceptualized in the latter way, which is inconsistent with the inverse U-shape observed in previous findings.

This paper is organized as follows: First, we introduce two of the most successful sampling-based models of human probability judgment, Probability Theory plus Noise (PT+N; Costello & Watts, 2014; 2016) and the Bayesian sampler (Zhu, et al. 2020). We then demonstrate, using a regression-based method, that both models will naturally account for the inverse U-shaped pattern found in data in a way that most deterministic models cannot, although they make qualitatively different predictions regarding the precise characteristics of this pattern; this is confirmed via model recovery. Lastly, by applying the same regression-based method to experimental data, we find that the results are consistent with a binomial sampling process with an adjustment according to a prior, as predicted by the Bayesian sampler.

Sampling-based models of human probability judgment.

Apart from varying stochastically, human probability judgment is subject to a number of biases. These biases are the focus of most models of human probability judgment, and they cannot be explained by sampling from underlying probabilities, which typically produce unbiased estimates. The most fundamental of these biases is conservatism: that people's probability judgments are usually less extreme than what would be expected (Costello & Watts, 2014; Erev et al., 1994; Hilbert, 2012; Kaufman et al., 1949).[1] Conservatism can, in turn, partly explain other biases, such as the conjunction and disjunction fallacies. A conjunction fallacy occurs when a conjunction between events is judged as more likely than either of the marginal events (such as judging the probability of a person being a bank teller and a feminist as more likely than being a bank teller) and, conversely, a disjunction fallacy occurs when a disjunction between events is judged as less likely than either of the marginal events (such as judging the probability of a person being a bank teller or a feminist as less likely than being a feminist; Bar-Hillel & Neter, 1993; Costello, 2009; Moro, 2009; Tversky & Kahneman, 1983). If one presumes that the conservatism bias is stronger for conjunctions and disjunctions than for single events, then conjunction and disjunction fallacies can occur (Costello & Watts, 2016; Zhu et al., 2020).

From the point of view of sampling accounts, where does conservatism come from? Two recent accounts make almost indistinguishable predictions but embody very different viewpoints on conservatism. One influential approach, the Probability Theory plus Noise model (PT+N; Costello & Watts, 2014; 2016), suggests that conservatism arises because the process of retrieving memories is corrupted by noise. So even if a person recalls, say, five days that were all rainy, there is a good chance they will erroneously recall one (or more) as dry, pulling their probability away from 1; and similarly, the noise will tend to push up

[1] Note that we focus specifically on conservatism in probability judgments, rather than conservatism in updates on probability estimates in light of new evidence (e.g., Peterson & Beach, 1967).

probabilities of low probability events. This account is similar to the noisy memory channels suggested by Hilbert (2012), but while Hilbert's theory includes various types of noise associated with various types of memory, PT+N focuses specifically on noisy retrieval of binary outcomes. More recently, Zhu et al. (2020) proposed a model in which conservatism arises not through noise, but as a result of Bayesian inference with small samples. Remarkably, the mean predictions of these models turn out to be identical. However, we will show that the two approaches differ regarding the relationship between the means and variances in probability judgments. Indeed, Bayesian models have a distinctive "signature" relationship between mean and variance, which is, we will see, empirically observed.

We first outline the two classes of theory in more detail. The PT+N model is based on the idea that for each sampled instance there is a certain probability (denoted *d*) that the outcome will be misread, so that an occurrence is read as a non-occurrence or vice versa. We will refer to this error as *retrieval noise*, to distinguish it from the *sampling noise* inherent in all sampling-based models with finite sample size. This is an important distinction because retrieval noise causes biased probability judgments while sampling noise does not. For an event *A*, the probability of reading an outcome as *A* according to the PT+N model is the weighted mean of the probabilities *P*(*A*) and *P*(*not*-*A*) weighted by *d*. Because the judgment of the PT+N model is based directly on the proportion of outcomes read as *A*, the average estimate is

$$E\left(\hat{P}(A)\right) = (1 - 2d)P(A) + d. \qquad 1$$

This means that whenever $d > 0$ the judgment will be regressed towards the middle of the scale, and also that the value of the *d* parameter will be directly related to the amount of bias that people show in their judgments (see Figure 1).

The Bayesian sampler, by contrast, does not assume that sampling is perturbed by retrieval noise, but rather that each judgment is adjusted by a prior on responses after sampling is complete and frequencies are tallied. The function of this prior is to compensate for the inherent uncertainty in sampling, by weighting the sampled proportion according to a distribution that is (presumably) representative of one's previous experiences and general knowledge. This prior on responses differs from the internal distributions from which outcomes are sampled in that it does not necessarily represent specific memories or simulated instances per se, but rather constitutes a conception of probability estimates in a more general and potentially more abstract sense; it is therefore insensitive to the specific details of the question at hand. In the Bayesian sampler, the prior is represented by a symmetric Beta distribution, which is commonly used as a prior over probabilities. The symmetric Beta distribution has a single parameter, b, that determines its shape: for $b > 1$ it is has a single peak at .5 and the probability is lowest at 0 and 1, for $b = 1$ it is a uniform distribution across the entire range, while for $b < 1$ it is a u-shape that peaks at 0 and 1 and the probability is lowest at .5.[2] If only a small number of sampled instances is used then such an adjustment will decrease average judgment error. For example, if one draws a red ball from an urn with an unknown proportion of red and blue balls, then it would be foolish to assume that the urn contained only red balls on that evidence alone; given no information on the proportion of red and blue balls (i.e., a uniform prior with $b = 1$) the estimate that will minimize average squared error is .67.

Notably, the Bayesian adjustment is a simple linear function of the sample size N and the prior parameter b. It also can be expressed in terms of the same parameter d as in the PT+N model and using the same equation (Costello & Watts, 2019; Zhu et al., 2020). In this case the d parameter does not represent retrieval noise but rather the influence of the symmetric prior as expressed by a Beta distribution, mitigated by the sample size. Zhu et al.

[2] Note that in many cases, including the article by Zhu et al. (2020), the parameter of the Beta distribution is represented by the letter β (or α and β) rather than b. We chose to change the notation to b in order to avoid confusion with the regression weight parameter, which is also traditionally denoted β.

(2020) show that the average estimate of the Bayesian sampler turns out to be mathematically identical to the average estimate of the PT+N model (see Equation 1), given the bridge conditions $d = \frac{b}{N+2b}$ and $(1 - 2d) = \frac{N}{N+2b}$. In both models d can therefore be considered a measure of bias, though in each model the bias has a different source (retrieval error and Bayesian adjustment, respectively). The bias expressed by the Bayesian sampler is also adaptive in the sense that *d* will decrease as the number of sampled instances increases and approach zero as sample size approaches infinity (see Figure 1).

These rather different interpretations of the source of bias in human probability judgment are difficult to disentangle. First, the predictions of mean probability judgments from the two models are identical for simple probabilities (e.g., the probability that a day will be rainy) and for conjunctions of two events (e.g., the probability that a day will be rainy and cold). The only judgments for which mean estimates appear different for the two models are for conditional probabilities (e.g., the probability that it will rain on a cold day). For these judgments, Zhu et al. (2020) found that the Bayesian sampler model outperformed the current version of the PT+N for conditional probabilities (Costello & Watts, 2016). But, as Zhu et al. (2020) note, the predictions of a slightly modified variant of the PT+N model can precisely match the predictions of the Bayesian sampler for conditional, as well as unconditional, probabilities. Thus, mean judgments alone do not provide the power to distinguish between retrieval noise and Bayesian adjustment due to using a prior over probabilities. In addition, using mean judgments alone only allows the overall bias *d* to be estimated, while the sample size *N* and prior parameter *b* are not identifiable.

A natural response to this state of affairs is to look at the individual judgments, because retrieval noise and the Bayesian sampler, as we discuss below, predict different distributions of responses. This is commonly done in a likelihood framework by calculating a model selection measure such as Bayes factors, AIC, BIC, etc. based on the likelihood of the model producing the observed data. Unfortunately for these two models this is not viable: each generally predicts that people will *never* use portions of the response scale. People do, at least on occasion, use any of the response values, and so the identity of the winning model will depend very heavily on the auxiliary assumptions about participant “response noise,” in particular the exact nature of the keypress and mouse movement errors that they make, rather than the theoretically interesting assumptions of each model (Acerbi et al., 2014).

In order to avoid these issues, Zhu et al. (2020) evaluated the fit of the models to individual judgments using a Wasserstein distance measure instead of likelihood. Wasserstein distance, often called earth mover’s distance, is the minimum amount of “effort” needed to transform one discrete distribution into another. This fit measure allowed for an evaluation of the goodness of fit of the discrete distributions predicted by the response-noise-free PT+N and Bayesian sampler models to the discrete distributions of responses made by participants. It also allowed the sample size N and prior parameter b to be estimated for the Bayesian sampler. While this measure has the potential to distinguish the models in theory, Zhu et al. (2020) found that it did not prove diagnostic in practice: the fits were somewhat ambiguous, and the parameter estimates appeared implausibly biased towards very large sample sizes for both models.

In this paper, we take a different approach. We show that it is possible to distinguish generally between retrieval noise and adjustment based on a prior by modeling the relationship between the mean and the variance of the judgments using linear regression. Recall that according to the PT+N model retrieval noise is an inherent part of the sampling process. By contrast, the adjustment described by the Bayesian sampler takes place after sampling is completed and proportions are tallied. It turns out that this distinction has different implications for the relationship between the mean and the variance of the probability judgments. In addition, linear regression can be used to extract parameter estimates of the sample size N and, for the Bayesian sampler, the parameter b of the symmetric prior distribution Beta(b, b). Focusing on the characteristic mean-variance signature in probability judgment tasks also highlights the challenge faced by deterministic models of probability judgments. Simply adding noise to the predictions of a deterministic model will not readily capture mean-variance signature observed in experimental data.

The mean/variance relationship as a linear regression

In sample-based models of human probability judgment, samples are distributed according to a binomial distribution, meaning that the variance of the estimate $\hat{P}(A)$ is dependent on the underlying probability[3] $P(A)$ as well as the sample size N, according to the equation

$$V\big(\hat{P}(A)\big) = \frac{1}{N} P(A)\big(1 - P(A)\big). \qquad 2$$

This formula illustrates that the variance of repeated judgments will be lower the larger the sample size N and be higher in the middle of the probability scale than at the edges of the scale. Although we typically do not have access to the underlying probabilities for probability estimates (which, in a sampling model, would be approached as the sample size increased to infinity), for repeated judgments we can generally use the mean of the estimates as a proxy. Thus, if we plot the variance of the estimates against the mean of the same estimates, we will observe an inverse U-shaped curve, the height of which will vary according to sample size (see Figure 1).

The retrieval noise of the PT+N model and the adjustment of the Bayesian sampler will both affect the variance of the estimates. However, because the PT+N model adjusts the probability of sampling A before outcomes are tallied, and the Bayesian sampler adjusts the estimate of $P(A)$ after outcomes are tallied, the two models will affect the variance in

[3] Note that “underlying probability” here refers to distributions in peoples’ minds, which might or might not correspond to the statistical properties of their environment.

different ways. Although both models predict the characteristic inverse U-shaped curve, and both predict that the height of the curve will vary depending on the size of N, the adjustment made by the Bayesian sampler will affect the relationship described in Equation 2 in a way that the PT+N models does not. This follows from the fact that the retrieval noise of the PT+N model only affects the variance in so far as it affects the probability of sampling an outcome, meaning that the relationship described in Equation 2 is retained, while a Bayesian adjustment made after sampling does not affect the probability of sampling an outcome but rather uses the sample size and prior (expressed as *d*) to adjust the probability judgment away from the relative frequencies, meaning that the relationship described in Equation 2 changes to include those additional parameters. This means that, when we plot the variance of the estimates against the means of the estimates, the retrieval noise of the PT+N model will not affect the shape of the curve, while the adjustment of the Bayesian sampler implies that the endpoints of the curve will be moved away from the extremes and towards the middle of the probability scale.

To get an intuition of why this is, imagine an event *Q* with probability *P*(*Q*) = 0. When the PT+N model estimates *P*(*Q*) it will potentially, depending on the value of *d*, mistakenly read a sample as *Q* rather than *not-Q*. This in turn implies that the probability of sampling *Q* is above zero, and therefore the variance of the judgments will also be above zero. In other words, the variance of the judgments has increased, but so has the mean, and therefore it has moved along the same curve. The Bayesian sampler, by contrast, will sample exactly zero occurrences of *Q* when *P*(*Q*) = 0, but will adjust the judgment to a value above zero, because it will compensate for small samples. In this case, because the probability of sampling *Q* is still zero, the variance of the judgment is also equal to zero, so the mean changes while the variance does not. In effect, the endpoints of the curve when using the Bayesian sampler will be adjusted in towards the middle (see Figure 1).

The mean-variance relationships described above can be expressed in terms of linear regression. As previously mentioned, we generally do not have access to the values of $P(A)$, but both the PT+N model and the Bayesian sampler allows us to use the expected values of the model predictions instead. This means that we can replace the underlying probability $P(A)$ in Equation 2 with the mean estimates $E\left(\hat{P}(A)\right)$. In the case of the PT+N model this is a simple substitution, while in the case of the Bayesian sampler we need to add an additional term to the equation. Putting the above considerations together, we can let $X = E\left(\hat{P}(A)\right)\left(1 - E\left(\hat{P}(A)\right)\right)$ be the independent variable, $\beta = \frac{1}{N}$ be the regression weight parameter, and $\alpha = \frac{1}{N}d(d-1)$, which is the additional term required by the Bayesian sampler, be the intercept (see Appendix A for the mathematical details), which means that we can express the variance of the estimates of the Bayesian sampler in terms of linear regression

$$V\left(\hat{P}(A)\right) = \alpha + \beta X. \qquad 3$$

This equation can be fitted to data with repeated judgments of different events, regardless of the relationship between events and without knowing the probabilities underlying the sampling process. Furthermore, the parameter estimates of the intercept α and the regression weight β can be used to in turn extract the parameters N, d, and b, by using the following equations:

$$N = \frac{1}{\beta} \qquad 4$$

$$d = \frac{1 - \sqrt{4N\alpha + 1}}{2} \qquad 5$$

$$b = \frac{Nd}{1-2d} \qquad 6$$

Note that the *d* in Equation 5 and 6 only pertains to the Bayesian sampler and not the PT+N model; as explained above, the *d* (i.e., the retrieval noise) of the PT+N model will not affect the relationship between mean judgments and variance. Because the variance of the estimates of the PT+N model can be expressed by the same regression as in Equation 3 but without the intercept, the presence of a (negative) intercept can be considered evidence in favor of the Bayesian sampler. In short, when fitting the regression described in Equation 3 to data, the PT+N model and the Bayesian sampler both predict a positive regression weight (because if $N > 0$ then $\frac{1}{N} > 0$), but the Bayesian sampler predicts a strictly negative intercept (because if $N > 0$ and $0 > d > .5$ then $\alpha = \frac{1}{N}d(d-1) < 0$) while the PT+N model predicts an intercept of zero. In both cases we can therefore estimate the sample size from the regression weight β (see Equation 4) and, if the Bayesian sampler is supported by the presence of a negative intercept, then we can also estimate the *d* parameter and the prior (see Equation 5 and 6).

This method has some notable advantages over a related method presented in Howe and Costello (2020) that fits a relationship between the *true objective probabilities* and the variance of estimates. Firstly, the mean-variance curve we describe is invariant to retrieval noise while the relationship between objective probabilities and variance is not, so our method provides a better test of the use of a prior. Secondly, because the present method does not require knowing participants' underlying probabilities it is applicable to a much wider range of experimental designs, though we assume that participant's underlying probabilities should hold constant with respect to the same query. Thirdly, because the present approach fits parameters with linear regression, we can take advantage of methodological advances in linear regression (e.g., pooling data across participants using random effects) as well as its fast and easy-to-use implementations (e.g., we have included a simple implementation of the model using the programming language R in the supplementary information).

The mean-variance signature of response errors

Importantly, most types of error or adjustments, such as retrieval error, typing errors, or misunderstanding the question, typically either increase total variance or do not affect total variance at all. We have already demonstrated that the linear regression described is invariant to retrieval error and it is easy to see that this holds for any type of noise that directly affect the probability of sampling an outcome, since changes in the value of variable $P(A)$ will not affect the relationship described in Equation 2. Typing errors or misunderstandings, on the other hand, both imply additional variability since they each represent a probabilistic change to the judgment; if we conceptualize typing errors and misunderstandings as independent random variables in their own right then they will each add to the total variance, since the variance of a sum of two independent random variables equals the sum of the variance. By contrast, neither the PT+N model nor the Bayesian sampler involve any additional random variables; the retrieval error of the PT+N model is subsumed by the sampling error and the Bayesian sampler does not include any random components beyond the sampling error. Note that this does not imply that the PT+N or the Bayesian sampler precludes other sources of variance, nor that a negative intercept implies that other sources of variance are not present; to the extent that they are, however, their only effect on the negative intercept implied by the Bayesian sampler would be to make it more difficult to detect. Therefore, a negative intercept is a strong indication of a Bayesian adjustment; to our knowledge there is no other model that consistently predicts this pattern.

There are of course many other cognitive models that can explain conservatism and other biases in human probability judgment. Although a formal comparison between all extant models of human probability judgment is beyond the scope of this paper (rewarding though such a study might be), it is nevertheless pertinent to determine what predictions other models would make concerning the mean-variance relationship, and hence the predicted results of the regression model discussed above. Critically, sampling-based models naturally account for the variability in human probability judgment by the stochasticity in the sampling process, while most other models (e.g., heuristics or averaging models) lack any inherent stochasticity and therefore can only account for the variability by assuming either that judgments are perturbed by the type of response noise described above or by adding a more loosely defined but mathematically equivalent "error term" to the model. Such deterministic models with an additive error term generally predict a positive intercept and a regression weight close to zero (see Appendix B for simulations demonstrating this using the Configural Weighted Average model; Nilsson et. al. 2009) and are therefore incompatible with the predictions made by both the PT+N model and the Bayesian sampler. We will return to a more in-depth treatment of these questions in the Discussion.

Model Recovery

As described above, the linear regression can be used to estimate the sample size *N*, as well as identify the presence of post-sampling adjustment and estimate the prior Beta(b, b) according to which judgments are adjusted. In previous research Zhu et al. (2020) estimated *N* and Beta(b, b) parameters using Wasserstein distance, but because overall fit for all models was better the larger the sample size, there is reason to suspect that this method results in a overestimation of *N*. To compare the accuracy of parameter estimates from the linear regression with estimates from Wasserstein distance, we performed model recovery using data with the same design and structure as in Experiment 1 in Zhu et al. (2020).

The simulated data for the model recovery consisted of 100 simulated participants and was created to match the empirical data as closely as possible. Therefore, each simulated participant was based on two sets of 20 unique queries, for a total of 40 unique queries. Each set of 20 queries were based on an event pair and, for each event pair, the 20 queries included judgments of the marginal events and their negations as well as all possible conjunctions, disjunctions, and conditionals. Each simulated participant was assigned a set of “true” probabilities for each marginal probability in each event pair drawn from the uniform distribution $U(0, 1)$. The true probabilities for each query were extrapolated from the marginal probabilities; for simplicity we assume independent events, so that $P(A|B) = \mathrm{P}(A)$. Each participant was also assigned a sample size N randomly drawn from the values {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 50, 100, 250}. These values were chosen to represent a range of both small and large sample sizes from previous literature, while also including steps representative of differences in scale (e.g., the step between $N = 5$ and $N = 6$ is of more relative importance than the step between $N = 105$ and $N = 106$). Lastly, each simulated participant was also assigned a Bayesian adjustment according to a prior parameter b that was randomly drawn from the values {0.2, 0.25, 0.33, 0.5, 1, 2, 3, 4, 5}. To create the simulated judgment data, three binomial samples with sample size N, representing three repetitions of each unique query, was drawn for each query and adjusted according to the Bayesian sampler with a prior Beta(b, b) (Zhu, et al. 2020), creating a dataset with a total of 120 judgments for each simulated participant.

Results indicate that estimates of N and b from the linear regression indeed have both less total error and less bias than estimates from Wasserstein distance, but that the estimate of d is comparable (see Figure 2 and Table 1).[4] We can conclude that Wasserstein distance is suitable for estimating the total amount of bias, but less so for estimates of sample size and prior.

[4] When recovering parameters, bounds are imposed so that $\alpha \leq 0$ and $\beta \geq 0$, in order to conform to the same conditions as for the Wasserstein modelling in Zhu et al. (2020). Without these bounds the results are broadly the same but there are occasional additional outliers (see Appendix C). These bounds are not used when recovering the number of statistically significant negative intercepts.

Table 1.

Mean Error and Mean Absolute Error for Recovered Parameter Estimates.

Parameter		**Linear regression**	**Wasserstein distance**
N	Mean error	.794	20.2
	Mean abs. error	7.11	20.3
b	Mean error	-.115	.709
	Mean abs. error	.729	.825
d	Mean error	-.007	-.011
	Mean abs. error	.017	.015

Because one of the most important aspects of the linear regression is its ability to distinguish Bayesian adjustment by the presence or absence of a negative intercept, we also created 100 simulated datasets each with 80 simulated participants (the average number of participants in our empirical datasets), generated in the same manner as above, and applied a mixed-effects model to each dataset with random slopes and intercepts for each participant. For each simulated dataset, we found a statistically significant (negative) intercept at p < .005. This procedure was then repeated but with simulated participants without a Bayesian adjustment, which resulted in no statistically significant intercepts for any of the simulated datasets. This was also confirmed using Bayesian regression; for all simulated datasets with a Bayesian adjustment we found a BF > 1,000 in favor of a model with intercept as opposed to without, and for all datasets without a Bayesian adjustment the same comparison resulted in a BF < .001. Although for individual participants negative intercepts close to zero may be difficult to detect (see Appendix C), we can conclude that, on an aggregate level, the linear regression will reliably detect a Bayesian adjustment if it is present.

Empirical Results

Having validated our method with simulated data, we now use it to distinguish between the Bayesian sampler and PT+N models on empirical data. Specifically, we applied the linear regression described above to data from two previous experiments (Zhu et al., 2020)[5] and two new datasets with the same experimental design.[6] The purpose of Experiment 3 was to replicate previous results, while also avoiding the possibility of schematic reasoning due to logical contradictions (e.g., warm and snowy weather) present to some degree in earlier experiments, and the purpose of Experiment 4 was to determine whether those same results also hold for future one-off events; the new experiments are therefore functionally identical to those in Zhu et al. (2020), except for the different nature of events in Experiment 4. In Experiment 1-3, participants made probability judgments of the format: “What is the probability that the weather is [X] on a random day in England?” Various weather events were used, and the queries included both marginal events, conditional events, conjunctions, and disjunctions. In Experiment 4, participants instead made probability judgments on future events, such as: “What is the probability that there will be an early UK general election?” Again, the queries included both marginal events, conditional events, conjunctions, and disjunctions, using the same structure as in Experiment 1-3 (see Appendix D for a description of the experimental design). Crucially, for all experiments, each individual query was repeated three times; the mean and variance for each participant’s judgment of each individual query was calculated from these three repetitions.

The linear regression described above was applied to each dataset as a mixed-effects model with participant as a random effect, in order to explore overall patterns while also allowing for between-subjects variability. These mixed-effects models were applied with

[5] Note that in Zhu et al. (2020), the event pair {*warm*, *snowy*} was excluded from Exp. 2, due to the fact that such highly dependent events can induce schematic reasoning if the participant consider, for example, the conjunction *warm and snowy* to be a logical contradiction. We found the results of the linear regression to be very similar in both cases and therefore chose to include also {*warm*, *snowy*} in the current analysis.

[6] Data for both new experiments is available at: https://osf.io/9kea6/.

random slopes as well as intercepts, in order to allow for between-subjects variability in sample size as well as prior, using an unstructured covariance matrix. The best fitting parameter values are summarized in Table 2 and plots of the best fitting function values are visualized in Figure 3. The results have three important implications. Firstly, we estimate that only a small number of instances were sampled for each judgment, as few as approximately three in some cases. This is in contrast to the notably higher medians of *N* (median N_{Exp1} = 10, median N_{Exp2} = 48) found by Zhu et al. (2020) using Wasserstein distance and, following the model recovery, we consider the results of the linear regression to be more reliable. Smaller samples are arguably more psychologically plausible as well; previous research has indicated that people only generate a small number of thoughts when coming to a decision (e.g., 3.6 on average; Weber et al., 2007), which is compatible with research on short-term memory capacity (Cowan, 2001). Additionally, it has been demonstrated that only very small samples are optimal when sampling is costly in time or effort (Vul et al., 2014).

Secondly, the linear regression confirms that judgments were indeed adjusted after the sampling process, as predicted by the Bayesian sampler. All four experiments showed a negative intercept (*p*-values for intercept: $p_{Exp1\text{-}3} < .001$, $p_{Exp4} = .005$; Bayes Factors in favor of the regression models with intercept compared to without intercept: $\text{BF}_{\text{Exp1}} > 1{,}000$, $\text{BF}_{\text{Exp2}} = 259$, $\text{BF}_{\text{Exp3}} > 1{,}000$, $\text{BF}_{\text{Exp4}} > 1{,}000$).[7] While the results in Zhu et al. (2020) did favor the Bayesian sampler overall, the fits were sometimes ambiguous, and the distinction relied primarily on the conditional probabilities. Here, by focusing the relationship between mean and variance, the linear regression clearly indicates that, on aggregate, a Bayesian adjustment did take place. This does not necessarily preclude other sources of bias, such as retrieval noise, but it does strongly confirm a key and distinctive prediction of the Bayesian sampler.

[7] The Bayesian mixed model regressions, as well as all other Bayesian regressions cited in this study, were performed using the R package BRMS v. 2.17.0, applied with a flat (proper) prior with bounds [-.25, .33] for the intercept and [0, 1] for the regression weight. Parameter estimates for the Bayesian regressions were practically identical to the frequentist regressions.

Thirdly, the adjustments suggest a U-shaped prior, because *b* < 1 for all experiments, with higher priors associated with extreme probabilities. Again, this contrasts with the findings by Zhu et al. (2020) and, again, we consider the results of the linear regression more reliable, since model recovery indicates that the estimates of the Wasserstein distance are positively biased. Interestingly, the estimated values are remarkably similar to the empirical prior that Zhu et al. (2020) calculated from the statistics of probability words in natural language observed by Stewart et al. (2006). In the latter study, the frequencies of when a range of probability related words and phrases were used to describe a probability were collected, and participants were then asked to judge the probability they associated with each phrase. When fitting a symmetric Beta distribution to the resulting distribution of values, the best fitting distribution was Beta(.27, .27), which is very similar to the results of the linear regression.

Table 2.
Best Mixed-Effects Model Parameter Estimates for Exp. 1-4.

Experiment	**Regression Parameters**		**Model Parameters**		
	α	β	N	b	d
Exp. 1*	-.026	.372	2.68	.239	.076
Exp. 2*	-.007	.150	6.66	.347	.047
Exp. 3**	-.024	.324	3.09	.293	.080
Exp. 4**	-.006	.156	6.43	.290	.041

*From Zhu, Sanborn, & Chater, 2020; **New dataset

Examining the random effects of the mixed-effects model also indicates very strong negative correlations between intercepts and regression weights, r_{Exp1} = -.811, 95% CI [-.884, -.700], r_{Exp2} = -1, 95% CI [-1, -1], r_{Exp3} = -.966, 95% CI [-.979, -.946], r_{Exp4} = -1, 95% CI [-1, -1] which in turn implies that bias is indeed adaptive (i.e., we observe more bias for smaller samples). Because intercepts and the regression weights are statistically related however,[8] it is more prudent to examine the extracted parameters.

To explore parameter correlations, we applied the linear regression to each participant individually, in order to avoid the random effects shrinking the individual estimates towards the mean. Results confirm that sample sizes are centered around relatively small values (median N_{Exp1} = 2.65, median N_{Exp2} = 8.25, median N_{Exp3} = 3.00, median N_{Exp4} = 5.91; see Figure 4). If bias, as measured by the parameter *d*, is adaptive, then we should find a negative correlation between *N* and *d*, and a positive correlation between *d* and *b*, because *d* is defined as $d = \frac{b}{N+2b}$. Conversely, we do not expect a correlation between *N* and *b*, because we have no a priori reason to expect sample size and the prior to be related. These predictions were all confirmed (see Table 3). While only a limited proportion of individual participants in the datasets had statistically significant intercepts (Exp. 1: 21/59, Exp. 2: 4/84, Exp. 3: 27/69 Exp 4: 24/111), this is to be expected with the current experimental design and parameter values (as demonstrated by model recovery in Appendix C). These effects were mirrored when counting the proportion of individual participants with BF > 3 in favor of a model with an intercept compared to without (Exp. 1: 14/59, Exp. 2: 3/84, Exp. 3: 9/69 Exp 4: 16/111). Nevertheless, one-sample t-tests could confirm that individual participants' intercepts were significantly negative in all Experiments (*p*-values: $p_{Exp1\text{-}4} < .001$; Bayes Factors in favor of intercepts below zero: $\text{BF}_{\text{Exp1}} > 1{,}000$, $\text{BF}_{\text{Exp2}} = 157$, $\text{BF}_{\text{Exp3}} > 1{,}000$, $\text{BF}_{\text{Exp4}} > 1{,}000$).

Table 3.
Correlation Coefficients Between Individual Parameters Estimates for Exp. 1-4.

Experiment	**Correlation**

[8] In a linear regression of the form $y = \alpha + \beta x$, intercept and slope are negatively correlated as long as $x > 0$. Simulations indicate that randomly generated N = unif(2, 10) and b = unif(.25, 1.25) will create a correlation of only $r = -.164$ however, so the observed correlation is still notably higher than what would be expected for chance alone.

	N vs. *d*	*N* vs. *b*	*d* vs. *b*
Exp. 1*	*r* = -.440 [-.625, -.207]	*r* = -.198 [-.432, .061]	*r* = .783 [.659, .866]
Exp. 2*	*r* = -.535 [-.673, -.363]	*r* = -.177 [-.377, .040]	*r* = .742 [.628, .825]
Exp. 3**	*r* = -.396 [-.579, -.176]	*r* = -.014 [-.250, .223]	*r* = .794 [.686, .868]
Exp. 4**	*r* = -.391 [-.538, -.221]	*r* = -.003 [-.189, .184]	*r* = .611 [.480, .716]

*From Zhu, Sanborn & Chater, 2020; **New dataset

Transparency and Openness

None of the work in this manuscript was preregistered. We have made the new data reported in this paper available at the Open Science Framework webpage: https://osf.io/9kea6/. R-code for applying the model is available as supplementary material.

Discussion

In this article we demonstrate how sampling-based models naturally explain the inverse U-shaped relationship between the mean and the variance of probability judgments, which deterministic models with a generic additive error term cannot account for. We further demonstrate how to model this relationship with a linear regression in order to determine whether the conservatism bias in human probability judgment is due to a Bayesian adjustment after sampling or arises from retrieval noise. We confirm through model recovery that this method will recover the relevant parameters. We apply the method to data from two published experiments and two new experiments, where we find that parameters are consistent with small samples and an adjustment after sampling, as predicted by the Bayesian sampler. It is worth noting that, although previous research has demonstrated that the Bayesian sampler fits data well (Zhu et al., 2020), this is the first direct demonstration of the presence of such an adjustment.

The parameters extracted from all three experiments are consistent with previous empirical data, in two important aspects: Firstly, the estimated number of samples are generally small. This is consistent with research on the limitations of working memory (Cowan, 2001; Miller, 1956; Weber et al., 2007) and the computational restraints of the brain (Gershman et al., 2015; Griffiths et al., 2015), as well as results indicating that only a very small number of samples is optimal when sampling takes time or effort (Vul et al. 2014). Additionally, it has been shown that some cognitive biases can be explained as a direct consequence of a small number of samples (Juslin et al., 2007).

Secondly, the parameters are consistent with a U-shaped prior, implying that most events are either very likely or very unlikely and that intermediate probabilities are more rarely encountered. This matches the distribution of probability words in natural language (Stewart et al., 2006), which tentatively suggests that the prior is based on the frequencies of probability concepts in day-to-day life. If this is indeed representative of general tendencies in people, then this might have interesting implications regarding how our usage of probability in everyday language affects the way we approach uncertainty, and vice versa.

Successful models will naturally explain data well, meaning that distinguishing two successful models only by their relative fit to data will be difficult. This is particularly the case with the PT+N model and the Bayesian sampler, since they make the exact same average predictions in many situations. The linear regression presented here is an alternative way of distinguishing between the models, by taking a conceptual difference between the two (i.e., the source of the conservatism effect) and exploring its effect on the variance of the data. Results indicate that the Bayesian sampler is consistent with data, insofar that we can confirm that estimates do indeed appear to be adjusted after sampling. Such an adjustment is necessary, as well as intuitively compelling, whenever sample sizes are small, meaning that the resulting bias is “adaptive” in the sense that it generally increases the accuracy of probability judgments at the cost of introducing inconsistencies such as conjunction fallacies. In a broader perspective, the relative adaptivity (or “rationality”) of the Bayesian sampler exists in an intersection between computational and Bayesian rationality, in that it assumes that, on the one hand, the human mind has limited computational capacity but, on the other hand, it compensates for these limitations in ways that will (on average) minimize the deviance between probability estimates and statistical reality (see Appendix C of Zhu et al., 2020). A recurring challenge of Bayesian models is that many Bayesian calculations are much too computationally complex for the human mind; the Bayesian sampler, by contrast, assumes that mind merely approximates Bayesian inference to the best of its capacity given the most effective tools at hand (e.g., sampling).

Although the linear regression validates the Bayesian sampler, responses could potentially be affected by other factors as well, such as errors in retrieving items from memory, which are central to the PT+N model. Indeed, the mechanisms of the PT+N model are not explicitly contradicted by the Bayesian sampler, and it is possible that both processes occur simultaneously, in which case a hybrid model might be called for. Because the mean-variance relationship we base our method on is invariant to retrieval noise, our results do not exclude this possibility. It should be noted however, that previous work on the PT+N model generally assumed a relatively low value of approximately $d = 0.1$ (Howe & Costello 2020; Costello & Watts, 2017), which might allow little room for retrieval noise after the Bayesian adjustment is accounted for. This also raises further questions regarding the source and function of the proposed retrieval noise and its relationship to the Bayesian interpretation. From the perspective of the PT+N account, although it is reasonable to assume that recall from long term memory is less than perfect, it appears odd that the same limitations should apply to mental simulation, as has been previously suggested in order to account for unique events (e.g., Costello & Watts, 2018a; Ludwin-Peery et al., 2020). From the perspective of the Bayesian sampler, on the other hand, it seems reasonable to assume that, if the Bayesian sampler compensates for small samples, then it should also adjust for errors in sampling, such as retrieval noise. A hybrid model of human probability judgment would need to take these points into account.

Further development of the Bayesian sampler will also need to relax the assumption that the samples drawn from each query are independent of one another. This will be necessary to explain effects of question order, e.g., how the percentage of individuals responding “yes” to the questions “Is Clinton honest?” and “Is Gore honest?” depends on the order in which the questions are asked, and which have been successfully accounted for by Quantum Cognition approaches (Wang et al., 2014). Sampling models using samples drawn independently for each query, such as the Bayesian sampler, do not account for these dependencies. While a model with full dependence (i.e., answering all queries with the same set of samples) would be deterministic and so would not predict the inverted U relationship between mean and variance, milder forms of dependence are possible. For example, an extension to the PT+N model that assumes samples can be primed by previous queries can account for key question order effects (Costello & Watts, 2018b). Extensions to the Bayesian sampler that assume local sampling instead of independent sampling are also under development (Zhu, et al., 2021) and should be evaluated against human-like question order effects.

When it comes to other models of human probability judgment, the characteristic inverse U-shaped relationship and the negative intercept constitutes empirical hurdles that models need to explain in order to fully account for the process. Defining judgment noise solely as a generic additive error term, as is the case in many models, is clearly inconsistent with the empirical mean-variance relationship. For example, some of the most successful types of cognitive models in the study of judgment and decision making are based on heuristics. Although these models are more often applied in the context of decision making under risk than probability judgment, the same basic principle generally applies: instead of performing a complicated estimation process to generate a probability judgment, a more accessible number is supplied using some simpler procedure. These models rarely make allowances for the stochasticity of human judgment, outside of the aforementioned generic error terms. Because these error terms are generally additive and independent of the value of the estimated probabilities, such models will neither predict the characteristic inverse U-shape nor a negative intercept when modeling the variance using the regression model; to the extent that they can be said to make any predictions at all, this prediction should be a positive intercept and a regression weight equal to zero.[9]

The same principle holds for averaging models such as the Configural Weighted Average model (CWA; Nilsson et al. 2009). This model assumes that, when judging the probability of a conjunction, people approximate a weighted average of the probabilities of the component events, with more weight given to the lower probability and vice versa. If one assumes that both the estimation of the component probabilities and the averaging process is perturbed by some (as yet undefined) type of error, as suggested by Nilsson et al., (2009), the result should be a positive intercept and a regression weight close to zero. We confirmed this principle with simulations using the CWA model as an example of a broader range of heuristic and averaging models (see Appendix B), though this analysis generalizes to other models that makes deterministic predictions. The outcomes show that an additive error term will produce a constant variance, with sharp downward turns at the edges if values are truncated at 0 and 1. In terms of the regression model, this implies a flat (or close to flat) curve with a positive intercept, which is not consistent with the inverse U-shape that is

[9] An exception to this would be if the response error is truncated at the edges of the scale, in which case it is possible that one would observe that variance might drop sharply at the very edges – Appendix B also shows that truncated errors do not predict a negative intercept except for extreme levels of noise.

experimentally observed. We must conclude that, although it is no doubt a useful strategy to predict variable behaviors in many other cases, in the context of human probability judgment the generic additive error term constitutes a mischaracterization of the underlying cognitive process.

Many other models make too few mechanistic assumptions regarding noise to make any kind of formal predictions. For example, the Inductive Confirmation model (Tentori et al., 2013) do not describe how probability judgments that are not directly associated with an implicit or explicit context are produced, and the Quantum Cognition account (Bruza et al., 2015; Pothos & Busemeyer, 2022) has so far not supplied a definitive account of noise and stochasticity in human cognition and therefore the model makes no predictions concerning the mean/variance relationship. However, the Bounded Log Odds model (BLO; Zhang et al., 2020) and other associated models working on the same principle (Khaw et al., 2021) constitute an interesting exception. Due to expressing noise as encoding variance on a linear log-odds scale, these models can produce the inverse U-shape and the negative intercept associated with the Bayesian sampler, though only for particular parameter values.[10] Because BLO does not make a strong prediction regarding the mean-variance relationship, but rather is consistent with different patterns depending on the configuration of parameters, we did not consider it further in connection to the main analysis.

It seems some kind of computational noise is necessary, and sampling-based models embody one form of computational noise. This fits well with past research that has indicated that Bayesian suboptimality in human inference cannot be explained by sensory or response noise alone, but necessitates some form of computational noise (Acerbi et al., 2014; Drugowitsch et al., 2016; Findling & Wyart, 2021; Stengård & van den Berg, 2019),

[10] More specifically, a standard deviation of approximately $\sigma > 0.8$ on the log-odds scale is required to mimic the magnitude of negative intercepts observed.

motivated by, for example, limited coding resources (Polanía et al., 2019). Sampling has been identified as a potential source of computational noise (Findling & Wyart, 2021) and Stengård and van den Berg (2019) in particular found that a simple form of sampling could partially explain the computational noise in their data. This is not to say that sampling-based models are necessarily the only models that could account for these effects; as previously mentioned, models based on a linear log-odds scale also have a form of computational noise that can in some cases mimic the same patterns, as might other models, given enough additional assumptions. The distinctive strength of sampling-based models, and the Bayesian sampler in particular, is that they predict these patterns as a natural consequence of the basic mechanisms of the model, rather than being reliant on specific parameters or auxiliary assumptions.

Although we here focus on the source of conservatism in human probability judgment and the distinction between two successful models in the area, we believe that the method we used has promise for wider application, particularly for estimating sample sizes for sampling-based accounts of judgments. Sampling-based models have generally been relatively non-committal regarding the number of sampled instances for each judgment, but there are indications that the human mind is limited to processing only a small number of samples (Cowan, 2001; Miller, 1956; Weber et al., 2007). A wider application of the linear regression presented here, or variants based on the same general idea, could go a long way to confirm or reject such claims.

Author Note

A preliminary version of our work was presented at the 42nd Annual Virtual Meeting of the Cognitive Science Society Conference and at the 2021 meeting of the Society for Mathematical Psychology, and a preprint of this work posted on PsyArXiv (https://psyarxiv.com/yuhaz), but otherwise none of the material included in this paper has been published or is under consideration elsewhere. Informed consent was obtained from all human subjects in the experiments described in this manuscript. None of the work in this manuscript was preregistered. We have made the new data reported in this paper available here: https://osf.io/9kea6/. Zhu, Sundh, and Sanborn were supported by a European Research Council consolidator grant [817492-SAMPLING].

Constraints on Generality

Our data samples were collected using the research participant pool at the University of Warwick and therefore largely consists of students at a UK university. Although we suspect that the basic principles (e.g., the inverse U-shaped relationship between mean judgments and variance) will generalize across different populations, more research should be performed to confirm this. Also, because the calibration and bias (e.g., overconfidence) of probability judgments have been shown to vary depending on both culture (Yates, 2010) and age (Prims & Moore, 2017), we can expect these differences to be expressed as differences in sample size and prior distribution (for the Bayesian sampler) or noise probability (for the PTN model). We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

References

Acerbi, L., Ma, W. J., & Vijayakumar, S. (2014, January). A Framework for Testing Identifiability of Bayesian Models of Perception. In *NIPS* (pp. 1026-1034).

Acerbi, L., Vijayakumar, S., & Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLoS computational biology, 10*(6), e1003661.

Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology, 65*(6), 1119.

Bruza, P. D., Wang, Z., & Busemeyer, J. R. (2015). Quantum cognition: a new theoretical approach to psychology. *Trends in cognitive sciences, 19*(7), 383-393.

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. *Trends in Cognitive Sciences*, *18*(10), 497-500.

Costello, F. (2009). Fallacies in probability judgments for conjunctions and disjunctions of everyday events. *Journal of Behavioral Decision Making, 22*(3), 235-251.

Costello, F., & Watts, P. (2014). Surprisingly rational: probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463.

Costello, F., & Watts, P. (2016). People’s conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, *89*, 106-133.

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.

Costello, F., & Watts, P. (2018a). Probability theory plus noise: Descriptive estimation and inferential judgment. *Topics in Cognitive Science*, *10*(1), 192-208.

Costello, F., & Watts, P. (2018b). Invariants in probabilistic reasoning. *Cognitive Psychology, 100*, 1-16.

Costello, F., & Watts, P. (2019). The rationality of illusory correlation. *Psychological Review*, *126*(3), 437.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87-114.

Drugowitsch, J., Wyart, V., Devauchelle, A. D., & Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron, 92*(6), 1398-1411.

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.

Findling, C., & Wyart, V. (2021). Computation noise in human learning and decision-making: origin, impact, function. *Current Opinion in Behavioral Sciences, 38*, 124-132.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*(6245), 273-278.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science, 7*(2), 217-229.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science, 21*(4), 263-268.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological Bulletin, 138*(2), 211.

Howe, R., & Costello, F. (2020). Random variation and systematic biases in probability estimation. *Cognitive Psychology, 123*, 101306.

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making, 10*(3), 189-209.

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review, 114*(3), 678.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology, 62*(4), 498-525.

Khaw, M. W., Stevens, L., & Woodford, M. (2021). Individual differences in the perception of probability. *PLoS computational biology, 17*(4), e1008871.

Kvam, P. D., & Busemeyer, J. R. (2020). A distributional and dynamic theory of pricing and preference. *Psychological Review, 127*(6), 1053–1078.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science, 31*(12), 1602-1611.

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

Moro, R. (2009). On the nature of the conjunction fallacy. *Synthese, 171*(1), 1-24.

Nilsson, H., Winman, A., Juslin, P., & Hansson, G. (2009). Linda is not a bearded lady: configural weighting and adding as the cause of extension errors. *Journal of Experimental Psychology: General, 138*(4), 517.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*(1), 29.

Polanía, R., Woodford, M., & Ruff, C. C. (2019). Efficient coding of subjective value. *Nature neuroscience, 22*(1), 134-142.

Pothos, E. M., & Busemeyer, J. R. (2022). Quantum cognition. *Annual review of psychology, 73*, 749-778.

Prims, J. P., & Moore, D. A. (2017). Overconfidence over the lifespan. *Judgment and decision making, 12*(1), 29-41.

Ren, X., Luo, H., & Zhang, H. (2021). Automatic and fast encoding of representational uncertainty underlies the distortion of relative frequency. *Journal of Neuroscience, 41*(16), 3692-3706.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Cciences*, *20*(12), 883-893.

Stengård, E., & Van den Berg, R. (2019). Imperfect Bayesian inference in visual perception. *PLoS computational biology, 15*(4), e1006465.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology, 53*(1), 1-26.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science, 38*(4), 599-637.

Wang, Z., Solloway, T., Shiffrin, R. M., & Busemeyer, J. R. (2014). Context effects produced by question orders reveal quantum nature of human judgments. *Proceedings of the National Academy of Sciences, 111*(26), 9431-9436.

Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice: A query-theory account. *Psychological Science, 18*(6), 516-523.

Yates, J. F. (2010). Culture and probability judgment. *Social and Personality Psychology Compass, 4*(3), 174-188.

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.

Zhu, J. Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*.

Zhu, J., Sundh, J., Spicer, J., Chater, N., & Sanborn, A. N. (2021, February 4). The Autocorrelated Bayesian Sampler: A Rational Process for Probability Judgments, Estimates, Confidence Intervals, Choices, Confidence Judgments, and Response Times. https://doi.org/10.31234/osf.io/3qxf7

Appendix A – Mathematical details

The variance of a probability estimate $\hat{P}(A)$, based on the proportion of occurrences *S* in *N* mental samples, is

$$\begin{aligned} V\big(\hat{P}(A)\big) &= V\left(\frac{S}{N}\right) \\ &= \frac{V(S)}{N^2} \\ &= \frac{NP(A)\big(1-P(A)\big)}{N^2} \\ &= \frac{P(A)\big(1-P(A)\big)}{N}. \end{aligned} \tag{A1}$$

This can also be written as

$$V\big(\hat{P}(A)\big) = \frac{1}{N}P(A)\big(1-P(A)\big), \tag{A2}$$

which, if we let the regression coefficient $\beta = \frac{1}{N}$ and independent variable *X* = *P*(*A*)(1 – P(*A*)), can be expressed as a linear model,

$$V\big(\hat{P}(A)\big) = \beta X. \tag{A3}$$

We generally do not have access to the values of *P*(*A*), but both the PT+N model and the Bayesian sampler can be rewritten to use the expected value of the model predictions instead. In this case, although we do not know the underlying probability *P*(*A*), we can substitute the mean response for each individual query. For both models $E\left(\hat{P}(A)\right) = (1-2d)P(A) + d$. In the PT+N model, *d* represents the probability of misreading a sampled instance, so that a positive outcome is read as a negative outcome or vice versa. In the Bayesian sampler however, *d* represents an adjustment according to a prior so that $d = \frac{b}{N+2b}$, where the prior distribution is defined as Beta(*b*, *b*). Because the PT+N model adjust the probabilities of sampling *A* before outcomes are tallied, and the Bayesian sampler adjusts the estimate of *P*(*A*) after outcomes are tallied, the variance is expressed in different ways.

For the PT+N model,

$$\begin{aligned} V\left(\hat{P}(A)\right) &= V\left(\frac{S}{N}\right) \\ &= \frac{V(S)}{N^2} \\ &= \frac{NP(\text{read as } A)\big(1 - P(\text{read as } A)\big)}{N^2} \\ &= \frac{P(\text{read as } A)\big(1 - P(\text{read as } A)\big)}{N} \\ &= \frac{\big((1-2d)P(A) + d\big)(1 - (1-2d)P(A) - d)}{N}. \end{aligned} \tag{A4}$$

By contrast, for the Bayesian sampler,

$$\begin{aligned} V\big(\hat{P}(A)\big) &= V\left((1-2d)\frac{S}{N} + d\right) \\ &= V(S)\left(\frac{1-2d}{N}\right)^2 \\ &= \frac{NP(A)\big(1-P(A)\big)(1-2d)^2}{N^2} \\ &= \frac{P(A)\big(1-P(A)\big)(1-2d)^2}{N}. \end{aligned} \tag{A5}$$

First, let us consider the PT+N model. Because *d* only affects the probability of sampling *A* (or rather, the probability of reading a sampled outcome as *A*, but these are mathematically equivalent) and because the estimate of the model is equal to the sampled proportion of *A*, we can assume that *P*(read as *A*) = $E\left(\hat{P}(A)\right)$. Therefore, we can ignore the last line of Eq. 4 and write,

$$\begin{aligned} V\left(\hat{P}(A)\right) &= \frac{E\left(\hat{P}(A)\right)\left(1 - E\left(\hat{P}(A)\right)\right)}{N} \\ &= \frac{1}{N}E\left(\hat{P}(A)\right)\left(1 - E\left(\hat{P}(A)\right)\right), \end{aligned} \tag{A6}$$

which gives us the linear model

$$V\left(\hat{P}(A)\right) = \beta X, \qquad \text{A7}$$

where $X = E\left(\hat{P}(A)\right)\left(1 - E\left(\hat{P}(A)\right)\right)$ and regression coefficient $\beta = \frac{1}{N}$. The parameter d is not part of the equation, because the degree of sampling noise in the PT+N model will affect the probability of sampling A but not the functional relationship of $P(A)$ to $V\left(\hat{P}(A)\right)$. Therefore, regardless of the value of d, the variance $V\left(\hat{P}(A)\right)$ will follow the same curve, defined by the parameter β.

The Bayesian sampler, on the other hand, assumes that the estimate is adjusted according to a prior *after* the sampling process is completed and the proportion of occurrences has been tallied. In the case of the PT+N model, the endpoints of the curve are located at $P(A) = 0$ and $P(A) = 1$ respectively, but this is no longer the case for the Bayesian sampler. Instead, the variance of the estimates is

$$\begin{aligned} V\left(\hat{P}(A)\right) &= \frac{P(A)\left(1 - P(A)\right)(1 - 2d)^2}{N} \\ &= \frac{1}{N} P(A)\left(1 - P(A)\right)(1 - 2d)^2. \end{aligned} \qquad \text{A9}$$

We can rewrite the equation in terms of the expected value of the responses, because the expected value of the responses is

$$E\left(\hat{P}(A)\right) = (1 - 2d)P(A) + d, \qquad \text{A10}$$

which we can rewrite as

$$P(A) = \frac{d - E\left(\hat{P}(A)\right)}{2d - 1}. \qquad \text{A11}$$

Inserting this into the above equation gives us

$$V\big(\hat{P}(A)\big) = \frac{1}{N}\left(\frac{d - E\left(\hat{P}(A)\right)}{2d-1}\right)\left(1 - \left(\frac{d - E\left(\hat{P}(A)\right)}{2d-1}\right)\right)(1-2d)^2, \qquad \text{A12}$$

which can be simplified to

$$V\big(\hat{P}(A)\big) = \frac{1}{N}E\left(\hat{P}(A)\right)\left(1 - E\left(\hat{P}(A)\right)\right) + \frac{1}{N}d(d-1). \qquad \text{A13}$$

This equation effectively gives us a linear model similar to the one described above, with $X = E\left(\hat{P}(A)\right)\left(1 - E\left(\hat{P}(A)\right)\right)$ and $\beta = \frac{1}{N}$ being exactly the same, plus the addition of the intercept $\alpha = \frac{1}{N}d(d-1)$. This does imply that the value of the intercept α is not wholly independent from the value of β, but because d is also dependent on the prior, α is still free to vary within certain bounds. This means that we can express the variance of the estimates of the Bayesian sampler as the linear model

$$V\big(\hat{P}(A)\big) = \alpha + \beta X. \qquad \text{A14}$$

These two linear models can then be compared to evaluate whether the pattern of variances is more consistent with the PT+N model or the Bayesian sampler.

Appendix B – Simulations of additive error

First, we randomly generated 10,000 pairs of marginal probabilities from a uniform distribution $U(0, 1)$. Each pair of probabilities was combined according to the Configural Weighted Average model (CWA; Nilsson et al., 2009), defined as

$$CWA = \gamma \min[P(A), P(B)] + (1 - \gamma) \max[P(A), P(B)], \qquad \text{B1}$$

where we set the weighting parameter $\gamma = 0.8$ (results are equivalent for $\gamma = 0.2$, which would be used for judging disjunctions). For each pair of probabilities, this process was repeated 10,000 times for each of noisy total estimates, noisy marginal probabilities, and noisy marginal probabilities as well as total estimates, and the mean estimate and mean variance was calculated. In each case, noise was defined as a Gaussian error term $\varepsilon \sim N(0, .1)$. This process was also repeated with values that were truncated at [0, 1] so that each value beyond 0 or 1 was set to the corresponding edge value (embodying the same assumptions used in the response error model of Juslin et al., 1997). Note that, for noisy total estimates, this is equivalent to any model that produces a deterministic probability estimate and adds a Gaussian error term, and therefore the results generalize to a broad variety of models. Plots of the mean and variance of simulated probability estimates are illustrated in Figure B1.

We can clearly see that the additive error term creates a uniform distribution of variance, with sharp downward curves at the very edges for the truncated values. If we apply the linear regression method introduced in the study, we can observe a nearly flat curve and a positive intercept (see Table B1). Although the curve estimated by the regression does show tendencies towards an inverse U-shape, it should be noted, firstly, that this will only appear if values are truncated and, even then, only if there is a significant proportion of values at the edges of the scale, and, secondly, that the regression weight is very small compared to what we observe in the empirical data. Most importantly, we see that this process will not replicate the negative intercept that is predicted by the Bayesian sampler, meaning that a negative intercept is strongly diagnostic. Of course, for the truncated values, increasing the level of noise will decrease the value of the positive intercept and at very large levels of noise ($\sigma > .3$) one might even start to observe a negative intercept. However, a model operating at this level of noise would have very questionable efficacy, so this seems unlikely to occur. By comparison, the median standard deviation observed in the experiments in this paper is ($\sigma \leq .1$) in all experiments, which is not associated with a negative intercept for truncated response noise.

Table B2.
Parameter estimates for the linear regression.

Parameter		**Intercept *α***	**Regression weight *β***
	Noisy total	.010	2.44×10^{-05}
Non-truncated values	Noisy marginals	.006	.001
	Noisy total & marginals	.016	.001
	Noisy total	.006	.016
Truncated values	Noisy marginals	.002	.018
	Noisy total & marginals	.006	.043

Appendix C – Model recovery

Model recovery without constraining parameter values

As demonstrated in Figure C1 and Table C1, running the linear regression with no bounds will result in a number of additional outliers, including several values that do not have a meaningful interpretation (N ≤ 0, *b* ≤ 0, *d* < 0). These outliers will contribute to larger bias and lower accuracy when compared to Wasserstein distance. Bounded parameters are therefore important in order to ensure reliable estimates of individual parameters.

Table C2.
Mean Error and Mean Absolute Error for Recovered Parameter Estimates with No Bounds.

Parameter		**Linear regression**	**Wasserstein distance**
N	Mean error	6.52	20.2
	Mean abs. error	12.8	20.3
b	Mean error	-1.73	.709
	Mean abs. error	2.35	.825
d	Mean error	-.015	-.011
	Mean abs. error	.025	.015

Intercepts for individual participants

In order to explore the proportion of statistically significant intercepts when analyzing individual participants, we created 100 simulated participants for each of *N* = {1, 2, 3, 4, 5, 10, 15, 20, 50, 100, 250} and each of *b* = {0.33, 1, 3} as well as without adjustment according to prior, then calculated the mean value of the intercept and the number of simulated participants with statistically significant intercept on the *p* < .05 level (see Table C3).

Table C3.

Mean Value of Intercepts and Number of Statistically Significant Intercepts Out of 100 Simulated Participants for Different Parameter Values (Number or Simulated Participants with BF > 3 for Model with Intercept Compared to Models Without Intercept in Parentheses)

N	No adjustment		*b* = 0.33		*b* = 1		*b* = 3	
	mean	stat. sig.	mean	stat. sig.	mean	stat. sig.	mean	stat. sig.
1	-5.29×10^{-17}	91 (99)	-0.24	100 (100)	-0.332	100 (100)	-0.368	100 (100)
2	-9.4×10^{-04}	0 (0)	-0.063	89 (60)	-0.112	99 (96)	-0.137	100 (100)
3	-8.29×10^{-04}	0 (0)	-0.031	50 (13)	-0.059	87 (84)	-0.082	97 (93)
4	-2.28×10^{-04}	0 (0)	-0.015	19 (4)	-0.039	79 (32)	-0.059	89 (69)
5	8.22×10^{-04}	0 (0)	-0.014	11 (4)	-0.024	54 (13)	-0.046	92 (68)
10	2.64×10^{-04}	0 (0)	-0.002	1 (0)	-0.008	16 (0)	-0.017	58 (16)
15	-5.93×10^{-04}	1 (0)	-0.001	0 (0)	-0.003	10 (0)	-0.007	37 (4)
20	-7.67×10^{-05}	0 (0)	-8.46×10^{-04}	0 (0)	-0.003	6 (0)	-0.005	25 (2)
50	1.62×10^{-05}	0 (0)	-3.44×10^{-04}	0 (0)	-6.6×10^{-04}	0 (0)	-0.001	3 (0)
100	-6.15×10^{-05}	0 (0)	1.22×10^{-05}	1 (0)	-2.08×10^{-04}	0 (0)	-2.36×10^{-04}	0 (0)
250	7.14×10^{-06}	0 (0)	-1.93×10^{-06}	2 (0)	-8.37×10^{-06}	0 (0)	-5.05×10^{-05}	1 (0)

We can see that the linear regression consistently outputs a negative intercept when an adjustment is present, but that for larger samples ($N > 10$) it is generally small enough that it will be difficult to detect with the current sizes of datasets. This is, to some extent, dependent on the nature of the prior, since a stronger prior implies larger adjustment and a larger (negative) intercept. It should be noted, however, that these results are based on an experimental paradigm mimicking that in Experiment 1 in Zhu et al. (2020), and that a higher-powered experiment would no doubt detect more statistically significant intercepts, though the mean values would presumably be the same.

It is also worth noting that the method detects a surprising number of statistically significant intercepts when no adjustment is made and the sample size is $N = 1$, although in this case the mean value is astronomically small. The reason for this is presumably that data can only take on two different values when the sample size is one, which can create illusory systematicity due to the lack of variance. This appears unlikely to happen in real data, but even if it did, the value of the intercept clearly demonstrate that no adjustment is actually taking place. Additionally, it would be very easy to detect if any participant was actually using only two different values in their responses.

Parameter variability

Although our modeling assumes constant sample size within subjects, it is of course possible that this varies to some extent. Although a full exploration of this possibility is beyond the scope of this paper, it is important to confirm that such variability would not affect the ability to distinguish between models. Therefore, we created 100 simulated datasets each with 80 simulated participants generated in the same manner as the model recovery in the main text, that is, two sets of 20 unique queries, making a total of 40 unique queries each repeated three times. Each participant was given a sample size parameter N randomly drawn from the values {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 50, 100, 250}, but for each data point the actual sample was drawn from a gamma distribution with a scale parameter $\theta = .5$ and a shape parameter set so that the mode of the distribution equals N for that participant. A Bayesian adjustment was applied with a parameter b randomly drawn from the values {0.2, 0.25, 0.33, 0.5, 1, 2, 3, 4, 5}. We then applied a mixed-effects model to each dataset with random slopes and intercepts for each participant. Again, for each simulated dataset with Bayesian adjustment, we found a statistically significant (negative) intercept at $p < .005$, and when the procedure was repeated with 100 simulated datasets without a Bayesian adjustment, we found no statistically significant intercepts.

Appendix D – Experimental design

All experiments were based on the basic probability judgment task introduced by Costello and Watts (2014, 2016), where participants were instructed to estimate the probability of a number of events. Responses were typed on a scale from 0 to 100.

Each set of queries was based on a number of event pairs (e.g., {*icy, frosty*}, see Table D1 for a comprehensive list). For each event pair, queries included the marginal events and their negations (e.g., *icy, not icy, frosty,* and *not frosty*) as well as all possible conjunctions, disjunctions, and conditionals between the marginal events and their negations (e.g., *icy and frosty, icy or frosty, icy given frosty, icy and not frosty* etc.). For each event pair, this made for a total of 20 unique queries. The total set of unique queries formed a block within which the presentation order was randomized for each participant, and each experiment consisted of three blocks, so that all participants responded to each unique query three times.

In Exp. 1-3, queries concerned the probability of various weather events occurring on a random day in England. Queries on marginal events, conjunctions, and disjunctions were expressed on the format "*What is the probability that the weather will be [some event] on a random day in England?*" and queries on conditional events were expressed on the format "*If the weather in England is [some marginal event] on a random day, what is the probability that weather will also be [another marginal event] on that same day?*"

In Exp. 4, queries concerned the probability of future events (e.g., the probability that Joe Biden will win a second term). Queries on marginal events, conjunctions, and disjunctions were expressed on the format "*What is the probability that [some event]?*" and queries on conditional events were expressed on the format "*If [some marginal event occurs], what is the probability that [another marginal event also occurs]?*" In order to avoid misinterpretation while keeping the wording of the actual queries relatively brief, the marginal events were all fully explained as part of the instructions (e.g., the instructions specified the marginal event as “that Joe Biden wins a second term as US president in the 2024 US election” while the query was expressed as “that Joe Biden will win a second term”). For conditional events, participants were specifically instructed that it does not matter which event occurs first or if one causes the other or not, instead they should assume that they (somehow) were given certain information about one event but not about the other.

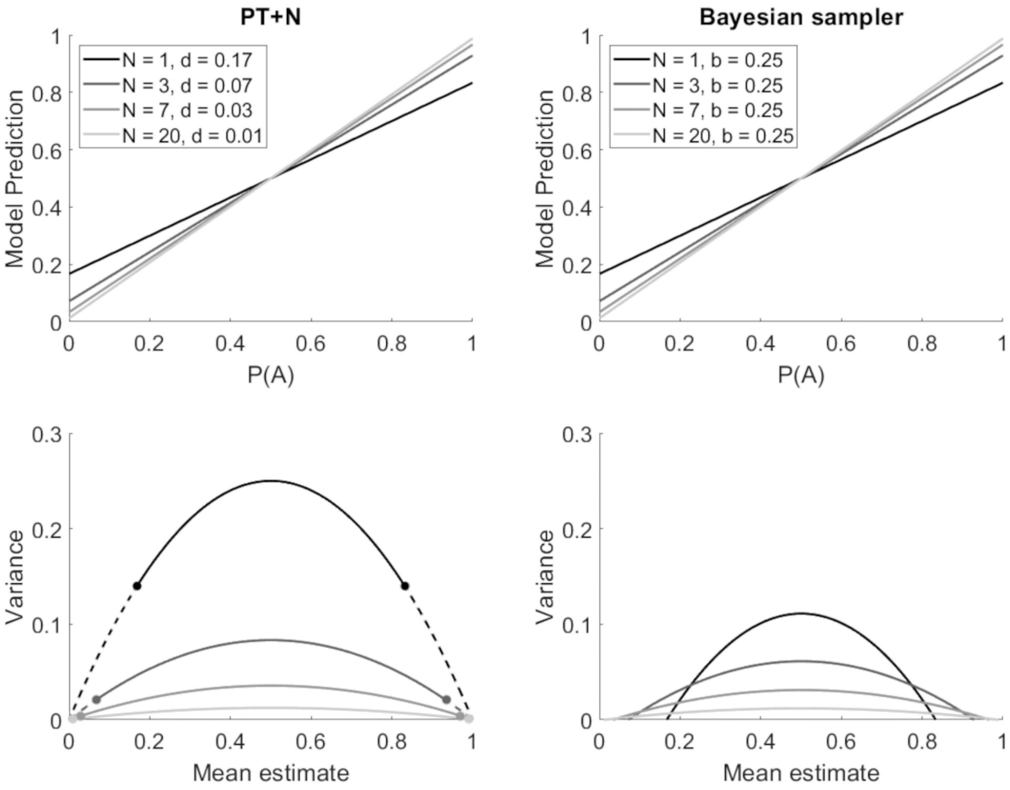
All experiments took approximately 30-40 minutes to complete. Participants for Exp. 1, 2, and 4 were recruited through the University of Warwick Student Research Experience Subject Panel and completed the experiment in exchange for course credit. Participants for Exp. 3 were recruited through the University of Warwick Research Participation System and were compensated with £3. Exp. 1-3 were administered in person at the psychology lab at the Department of Psychology, University of Warwick. Exp. 4 was administered online between 26 April and 7 May 2021, meaning that the events in the queries were still well in the future. See Table D1 for details.

Table D1.

Number of Participants Tested and Event Pairs for All Experiments

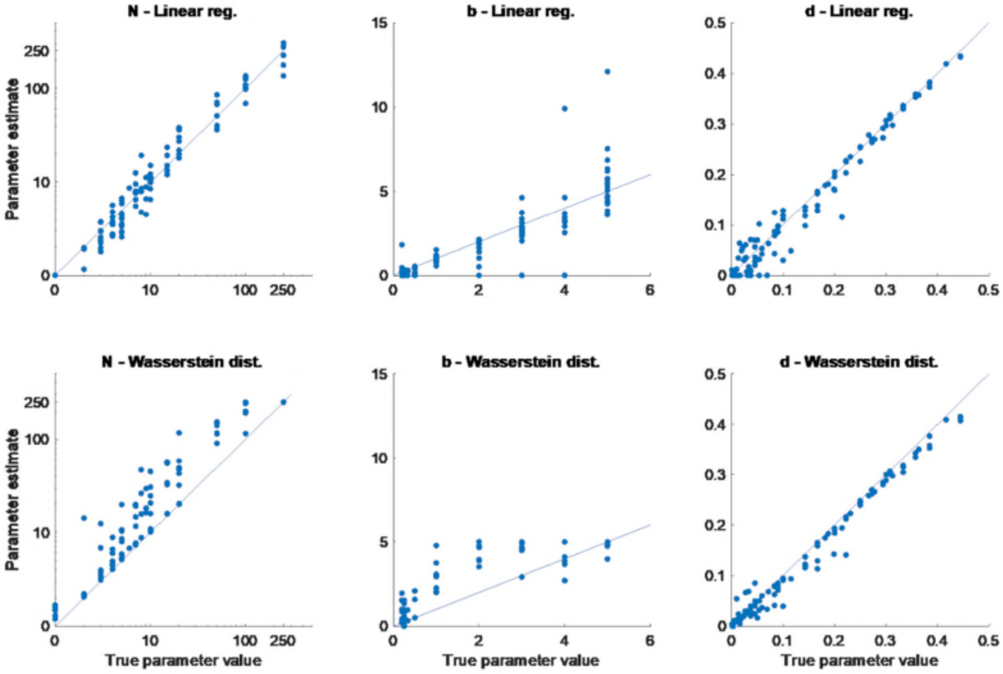
	No. participants	**Event pairs**
Exp. 1	59	{icy, frosty} {normal, typical}
Exp. 2	84	{cold, rainy} {windy, cloudy} {warm, snowy}
Exp. 3	69	{snowy, stormy} {thundery, humid} {mild, foggy}
Exp. 4	111	{Joe Biden winning a second term, the world reaching the 2050 climate goals} {early UK election, the UK economy recovering within current year}

Figure 1.



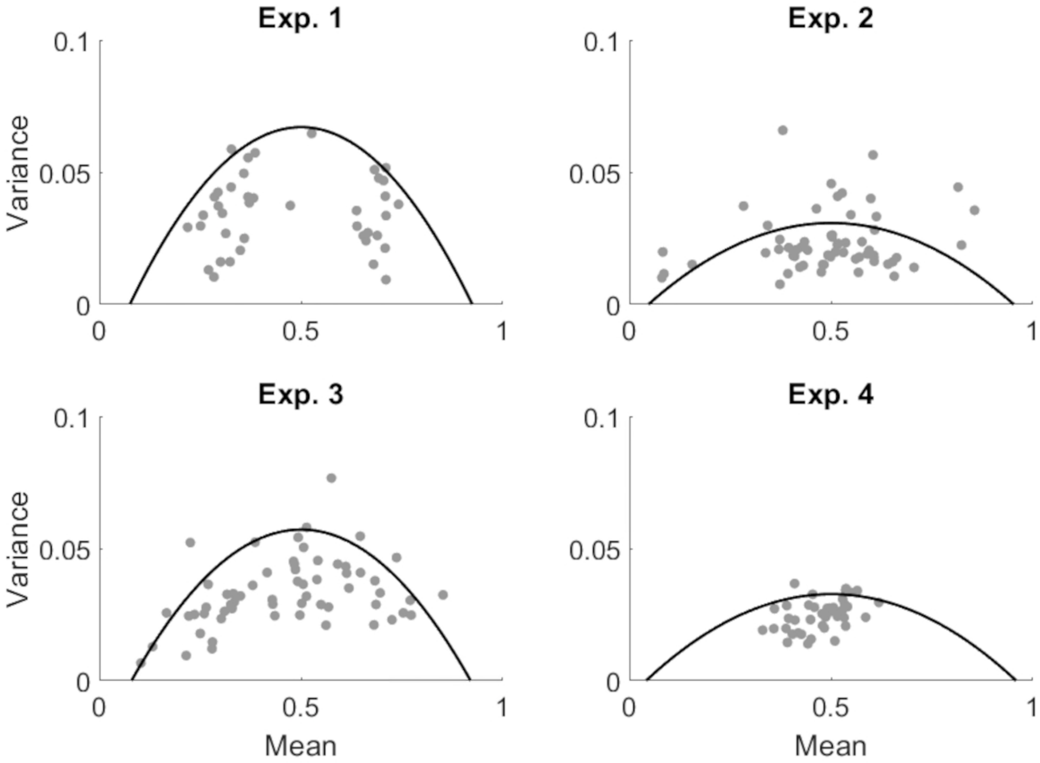
Note. Upper panels show the underlying probability (x-axis) and the corresponding predicted model estimation (y-axis) for different levels of bias and sample sizes. Lower panels show the mean model estimate (x-axis) and the corresponding variance for the same levels of bias and the same sample sizes. As demonstrated, the same levels of bias and sample sizes are associated with different relationships between mean estimate and variance for the PT+N model (left side panels) and the Bayesian sampler (right side panels). Note that, for the lower left panel, the predicted mean estimates of the PT+N model for *d* > 0 will be strictly above zero, but because it is important to demonstrate that the curve intersects the y-axis at zero we have chosen to retain the curve (dashed) for the rest of the interval.

Figure 2.



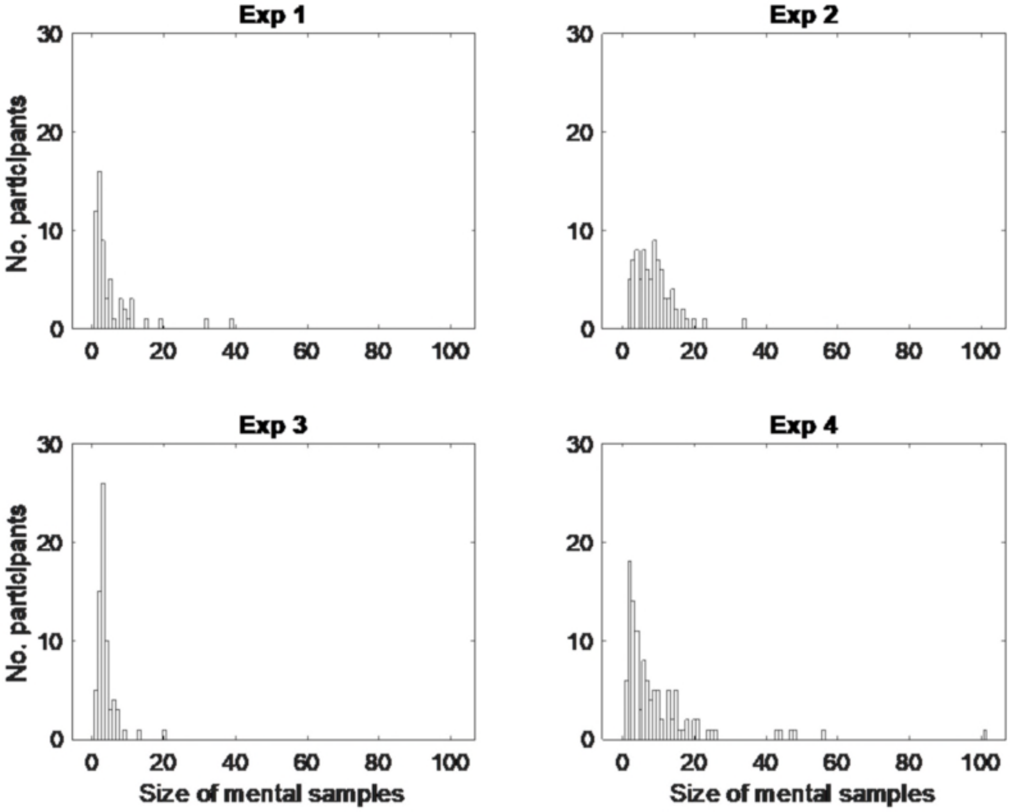
Note. Scatterplots of the true parameter values (x-axis) and the model estimates (y-axis) for the linear regression (upper panels) and the Wasserstein distance (lower panels). Note that, for readability, the plots of the N parameter (leftmost panels) are shown in log scale.

Figure 3.



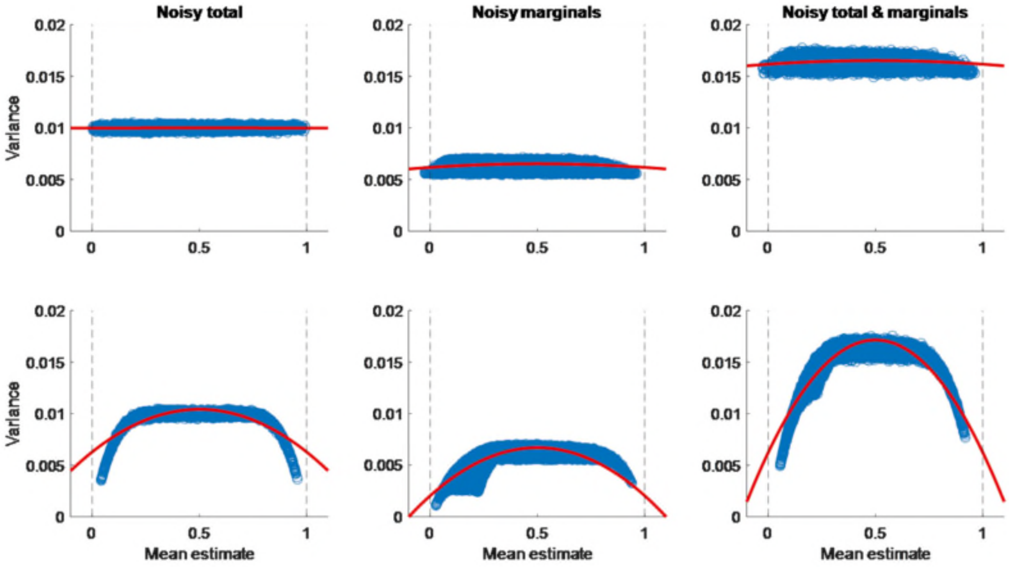
Note. Best fitting regression lines for the mixed-effects linear models for each experiment. Data points represent the mean estimate and mean variance of each individual query.

Figure 4.



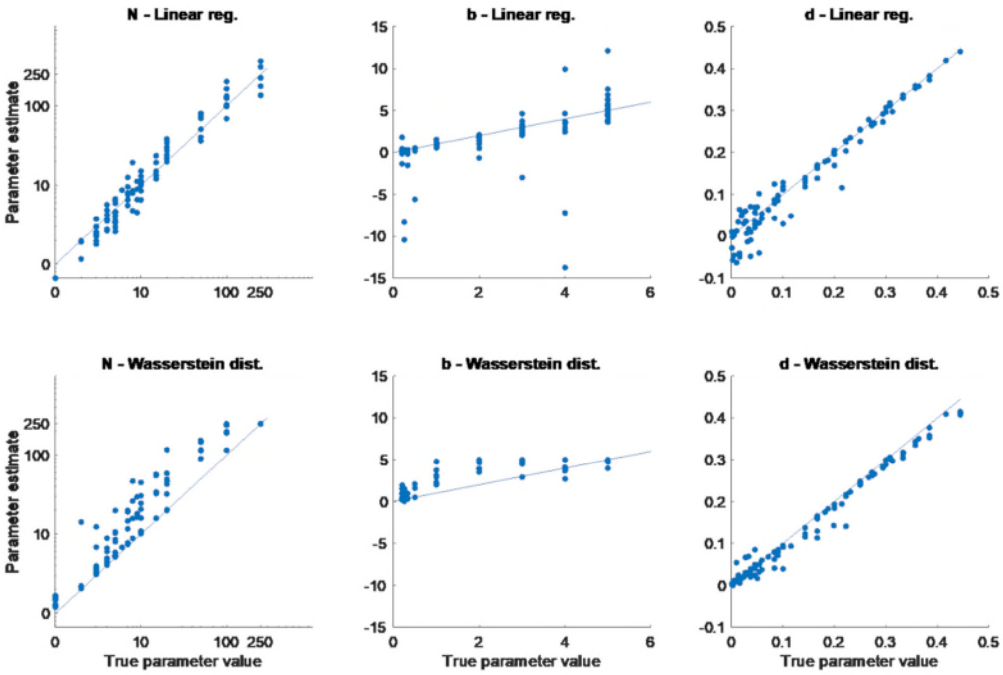
Note. Histograms of the distributions of fitted sizes of mental samples in all experiments.

Figure B1.



Note. Scatterplots of the mean (x-axis) and the variance (y-axis) of simulated probability estimates for different added noise with non-truncated (upper panels) and truncated (lower panels) values. The red lines represent the best fit of the linear regression model.

Figure C1.



Note. Scatterplots of the true parameter values (x-axis) and the model estimates (y-axis) for the linear regression (upper panels) and the Wasserstein distance (lower panels). Note that, for readability, the plots of the N parameter (leftmost panels) are shown in log scale, and one extreme outlier with linear regression estimates *N* = 412, *b* = -107, and *d* = -.542 is excluded from all panels.