

Base rate neglect and conservatism in probabilistic reasoning: Insights from eliciting full distributions

Piers Douglas Lionel Howe¹, Andrew Perfors¹, Bradley Walker²,
Yoshihisa Kashima¹, and Nicolas Fay²

¹School of Psychological Sciences, University of Melbourne

²School of Psychological Science, University of Western Australia

Abstract

Bayesian statistics offers a normative description for how a person should combine their original beliefs (i.e., their priors) in light of new evidence (i.e., the likelihood). Previous research suggests that people tend to under-weight both their prior (base rate neglect) and the likelihood (conservatism), although this varies by individual and situation. Yet this work generally elicits people's knowledge as single point estimates (e.g., x has 5% probability of occurring) rather than as a full distribution. Here we demonstrate the utility of eliciting and fitting full distributions when studying these questions. Across three experiments, we found substantial variation in the extent to which people showed base rate neglect and conservatism, which our method allowed us to measure for the first time simultaneously at the level of the individual. We found that while most people tended to disregard the base rate, they did so less when the prior was made explicit. Although many individuals were conservative, there was no apparent systematic relationship between base rate neglect and conservatism within individuals. We suggest that this method shows great potential for studying human probabilistic reasoning.

Keywords: Bayesian reasoning, representation, probability, belief integration, priors, optimality, base rate neglect, conservatism

Piers Howe  <https://orcid.org/0000-0001-6171-1381> Correspondence concerning this article should be addressed to Piers Howe, School of Psychological Sciences, University of Melbourne, Melbourne, VIC 3010. E-mail: pdhowe@unimelb.edu.au

Introduction

Bayes' theorem offers a normative account about how beliefs should be updated in light of new data. According to it, the probability of a belief or hypothesis H conditional on data D is:

$$P(H|D) \propto P(D|H)P(H) \quad (1)$$

where the likelihood $P(D|H)$ is the probability of the data given the hypothesis H , and the prior $P(H)$ reflects the degree of belief in the hypothesis before seeing the data. Across a wide variety of domains, Bayesian models have emerged as a powerful tool for understanding human cognition. One useful aspect of such models is that they provide a normative standard against which human cognition and decision making can be compared. This approach has been applied successfully in a wide variety of domains including concept learning (Kemp, 2012; Sanborn et al., 2010), causal inference (Lucas & Griffiths, 2010), motor control (Wolpert, 2009), and perception (Vincent, 2015).

Despite the success of the Bayesian approach, and though people in the aggregate sometimes appear to behave qualitatively in accordance with Bayesian reasoning (Griffiths et al., 2010), there is strong evidence that *individuals* usually do not. People tend not to update their beliefs exactly in accordance with Bayes' theorem, either by underweighting the prior (Kahneman & Tversky, 1973) or by conservatively updating the likelihood (Phillips & Edwards, 1966) or both (Benjamin et al., 2019). *Base rate neglect* occurs when people discount information about prior probabilities when updating their beliefs. It has been replicated in field settings and hypothetical scenarios (e.g., Bar-Hillel, 1980; Kahneman & Tversky, 1973; Kennedy et al., 1997; Tribe, 1971) as well as lab experiments involving samples of abstract objects such as sampling balls from urns (e.g., Grether, 1980; Griffin & Tversky, 1992). Interestingly, in addition to underweighting the prior, people also often underweight the likelihood: that is, they fail to update their beliefs as strongly as Bayes' theorem predicts. This phenomenon, known as *conservatism*, has also been widely replicated across a variety of situations (Corner et al., 2010; Grether, 1992; Hammerton, 1973; Holt & Smith, 2009; Peterson & Miller, 1965; Phillips & Edwards, 1966; Slovic & Lichtenstein, 1971).

To some extent, base rate neglect and conservatism cannot easily be separated. Assuming a point prior hypothesis and a single data point, in fact, it is mathematically impossible to identify whether the prior or the likelihood is responsible for a particular pattern of inference: a weaker inference than expected could reflect either conservative updating or stronger priors than were assumed, while a stronger inference than expected could reflect either weaker priors or overweighting the likelihood. Most research exploring base rate neglect and conservatism does not disentangle the effects of priors and likelihoods, and those studies that do focus on aggregate behaviour (see Benjamin et al., 2019, for a review). As a result, little is known about how conservatism and base rate neglect co-occur within the same individual. More problematically, as Mandel (2014) points out, people's priors are typically not measured at all; it is instead assumed that they correspond to the given base rate. However, if they do not – for instance, if participants are suspicious about the accuracy of the base rate or they represent it with some fuzziness in memory – this could look like conservative updating or base rate neglect when it is not.

Even those studies that explicitly measure people's priors are somewhat lacking, since virtually all of them elicit priors (and posteriors) as point estimates rather than full distributions (see, e.g., Benjamin et al., 2019; Mandel, 2014; Savage, 1971; Wallsten & Budescu, 1983, for overviews and discussion). This matters because, as illustrated in Figure 1, distributional shape plays an important role in belief updating: even perfect Bayesian reasoners whose priors have the exact same expected value may draw different conclusions if their priors have different distributional shapes. Thus, determining whether people update their beliefs in accordance with Bayes' theorem depends heavily on obtaining an accurate measure of the full distribution of prior beliefs.

Of course, this is only relevant if people actually *do* represent probabilities as distributions, at least implicitly. It is generally assumed that this is the case, as described by Wallsten and Budescu (1983) when discussing the measurement of subjective probabilities: "Upon being asked to evaluate the probability of an outcome, a person will search his or her memory for relevant knowledge, combine it with the information at hand, and (presumably) provide the best judgment possible... If the same situation were replicated a large number of times, and if the person had no memory of his or her previous judgments, the encoded probabilities, x , would give rise to a distribution for that particular individual" (p 153). This view reflects considerable (if often implicit) agreement; even those who suggest that people make specific inferences on the basis of samples rather than full distributions assume that the underlying representation from which the samples are generated is a distribution (e.g., Lieder & Griffiths, 2020; Mozer et al., 2008; Vul et al., 2009; Vul & Pashler, 2008). Indeed, there is a rich literature about how best to elicit and measure full belief distributions (see Schlag et al., 2015, for a review). This literature, which was developed in applied contexts such as political science and expert elicitation, has rarely been used in research on Bayesian belief updating and provides the methodology that we employ here.

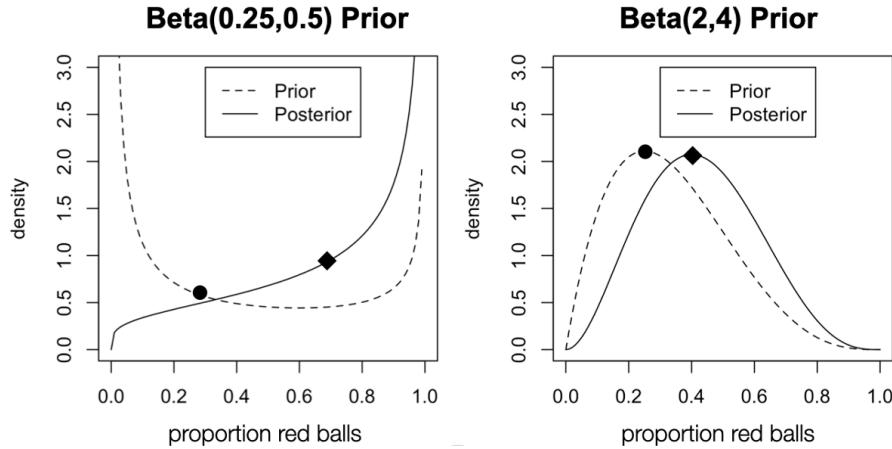
Our aim was to investigate the extent to which people demonstrate base rate neglect and/or conservatism in a simple probability task when their responses are full probability distributions rather than single estimates. Uniquely, we model not only the aggregate behaviour of all of our participants, but also each individual person's inferences based on their reported prior and posterior distributions. For each individual we are able to simultaneously estimate their degree of base rate neglect and their degree of conservatism based on their data from a single task. We elicited the prior and posterior distributions by asking people to draw histograms in an online visual interface involving bars; as Goldstein and Rothschild (2014) show, this method is intuitive to lay people and results in more accurate estimates than typical verbal reports.

The probability task consists of a game, common in this literature, known as the urn problem (Corner et al., 2010; Johnson & Kotz, 1977; Peterson & Miller, 1965). In a typical version of this game, participants are asked to imagine a container like an urn containing two different types of object (e.g., red and blue chips). Objects are drawn from the container and are revealed to the participant sequentially. Based on this information, people are asked to estimate the overall proportion of one of the types of objects (e.g., red chips) in the container.¹

¹Although the urn problem is traditionally used to study probabilistic reasoning, technically participants estimate

Figure 1

The importance of distributional shape in Bayesian updating



Note. The reasoners in both panels are perfect Bayesians who are estimating the probability of observing a certain outcome, such as pulling a red ball from an urn. Both have priors with the same expected value (single dots such that $P(\text{red}) = \frac{1}{3}$) but different full prior distributions (dotted lines). The prior on the left, Beta(0.25,0.5), reflects initial beliefs that the urn has either mostly red balls or mostly blue balls (probably mostly blue). The one on the right, Beta(2,4), reflects the belief that there are slightly more blue balls. This difference in distributional shape has a strong effect on the posterior distributions that are inferred after seeing a single new data point corresponding to one red ball. Not only do the posteriors have different distributional shapes, the expected values (diamonds) are also different: the one on the left has an expected value of $P(\text{red}) = 0.74$ and the one on the right $P(\text{red}) = 0.43$. This shows that fully capturing Bayesian updating requires getting the shape of the underlying distribution right, not just accurately measuring the point estimate of the expected value of the prior.

We report three experiments. In Experiment 1, we presented people with the urn problem but elicited their priors and posteriors as distributions rather than single estimates by having them draw histograms. We had two main goals in doing this: to establish what people actually assume if the prior is left unspecified, and to determine to what extent each person's reasoning was well-captured by Bayes' theorem using their stated prior. Our findings suggested that people showed substantial individual differences in their reported priors as well as how closely they followed the predictions of Bayes' theorem. That said, the majority demonstrated strong base rate neglect, with most people completely or almost completely disregarding their stated priors. They also showed a moderate degree of conservatism, updating their beliefs somewhat less than a fully Bayesian reasoner would, with no readily apparent systematic relationship between the two. We followed up in Experiment 2 by presenting people with explicit information about a stronger prior distribution in

proportions, not probabilities. However, in the urn problem they are equivalent; because all chips have an equal chance of being drawn, the probability that a red chip is drawn next is equal to the proportion of chips in the urn that are red. The advantage in explaining the task in terms of proportions rather than in terms of probabilities is that participants typically find the former formulation easier to understand (Gigerenzer & Hoffrage, 1995).

order to determine whether this changed the extent to which they incorporated it. Most participants still showed some conservatism and base rate neglect, although less strongly. To ensure that these results were not due to the particular prior used in that experiment, Experiment 3 used a different prior – the prior that, in the aggregate, people assume when not explicitly given a prior (as determined by Experiment 1). Experiment 3 confirmed that explicitly giving participants a prior caused them to neglect the base rate less than when they were required to infer the prior for themselves.²

Experiment 1

Method

According to Bayes' theorem, the degree to which a person's prior influences their posterior is determined by the amount of data they see: the more data, the more the posterior is shaped by the likelihood rather than the prior. We therefore studied scenarios which varied in the amount of evidence provided: some participants were shown just five chips drawn from the urn before they were asked to state their posterior, while others were shown five chips and then allowed to draw as many additional chips as they desired.

The MAIN condition of this experiment involved a within-subjects design that combined both scenarios. In it, participants reported three distributions: their prior, their posterior after seeing five chips, and their posterior after as many additional chips as they wanted. In order to ensure that the act of eliciting multiple distributions did not change how participants reported their estimates, we also ran two control conditions. In the ONLYFIVE condition people did not report a prior and reported only one posterior, after five chips. In the ONLYUNLIMITED condition participants did not report a prior and reported only one posterior, this time after viewing as many chips as they wanted.

Participants

452 people (249 male, 201 female, 2 non-binary) were recruited via Prolific Academic and paid 60 British pence. Mean age was 31 years old. Ninety were excluded because they failed the bot check (see below) or did not adjust any bars when estimating distributions. All participants gave informed consent and all three experiments in this paper were approved by the University of Melbourne School of Psychological Sciences Human Ethics Advisory Group (ID: 1544692).

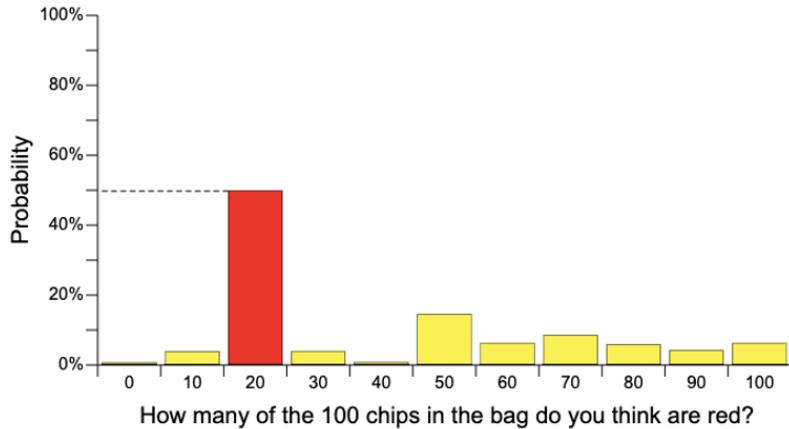
Materials

In all conditions, participants were shown an image of a bag that they were told contained red and blue chips. They were asked to provide their probability distributions by adjusting sliders corresponding to bars of a histogram, as shown in Figure 2. The first bar represented the participant's estimate of the probability that 0% of the chips in the bag were red, the second that 10% were red, and so on, with the final slider representing their estimate of the probability that 100% of the chips were red. The sliders were initialised randomly and constrained so that the total probability added

²Data and analysis code for all experiments can be found at <https://github.com/perfors/probability/>

Figure 2

Experiment 1: Elicitation of participants' probability distributions



Adjust the height of the bars on the graph to show how many of the bag's 100 chips you think are red. For example, if you think there is a 50% chance that 20 of the chips are red, drag the 20 bar up to 50% probability. You can raise or lower as many bars as you want, but note that they will automatically adjust to add up to 100% probability in total. When you are done, click "Submit".

If you are using a mouse, you should be able to drag the top of each bar. Otherwise, click [here](#) to use a different set of controls.

Note. A screenshot showing the methodology we used (in all experiments) for participants to report their probability distributions, similar to that of Goldstein and Rothschild (2014). People clicked on each bar to adjust its height. Clicking on a bar temporarily changed its colour to red. The different set of controls mentioned in the screenshot were a series of up and down buttons participants could press to adjust each slider. All probability distributions were constrained to sum to 100%.

up to 100%. In this way, by varying the position of the sliders, people could draw their probability distributions. When they were satisfied they pressed the submit button to continue.

Procedure

Bot check. All participants were initially asked a series of four multiple-choice questions to determine that they were human with adequate English abilities and not a bot. These questions posed analogies of the form “Mother is to daughter as father is to...” (in this example, the correct answer is “son”). Providing an incorrect answer to any of these questions counted as failing the bot check; data from these participants were not analysed. Following the bot check, instructions were presented, demographic information was collected, and people were allocated randomly to one of three conditions.

ONLYFIVE condition. Participants in this condition ($N = 126$) were shown an image of a bag that they were told contained red and blue chips. Five chips (four red and one blue) were then

drawn from the bag and presented to each participant one at a time in a random order. Participants were asked to report their estimate of the proportion of red chips in the bag using the histogram visualisation tool shown in Figure 2.

ONLYUNLIMITED condition. This condition was identical to the ONLYFIVE condition, except after the first five chips were presented, instead of reporting their posterior participants ($N = 129$) were given the option of drawing an additional chip. If they chose to draw one, after a delay of one second they were informed of the colour of the chip and given the option to draw another. This process could be repeated as many times as the participant desired. For the first five chips, four were red and one was blue, but the position of the blue chip in the sequence was randomised between participants. In every additional sequence of five chips, the pattern repeated: four chips were always red and one was always blue, but the position of the blue chip was randomised. After the participant was satisfied that they had drawn enough chips, they were asked to report their estimate of the proportion of red chips using the histogram visualisation tool shown in Figure 2.

MAIN condition. This condition was identical to the previous two except that each person ($N = 107$) was asked to estimate the distribution three times: once before being shown any chips, once after being shown five chips, and finally after having had the opportunity to view as many additional chips as they desired. Thus, each participant estimated one prior probability distribution and two posterior probability distributions, one after five chips and one after an unlimited number.

Modelling

Our research questions required determining the extent to which people under-weighted their prior and/or likelihood when reasoning about what chips they expected to see. We thus modelled participants as Bayesian reasoners who made inferences according to the following equation:

$$P(x|n_r, n_b) \propto P(n_r, n_b|x)P(x) \quad (2)$$

where x is their belief or hypothesis about the proportion of chips in the bag that are red and n_r and n_b represent the observed data (i.e., the number of chips that were drawn from the bag that were red and blue respectively). Thus, $P(x|n_r, n_b)$ is the posterior and $P(x)$ is the prior.

Prior. We captured each participant's level of base rate neglect with a parameter β which is a weighted average of the participant's stated prior, φ , and a uniform distribution U , as in Equation 3. The value for β is a constant ranging from 0 to 1, with $\beta = 0$ indicating that the stated prior was ignored entirely when calculating the posterior (i.e., complete base rate neglect) and $\beta = 1$ indicating that the prior was weighted appropriately (i.e., no base rate neglect at all).

$$P(x) = \beta\varphi + (1 - \beta)U \quad (3)$$

Likelihood. In an urn problem such as ours, $P(n_r, n_b|x)$ is captured according to a binomial likelihood function. In order to capture the extent to which each participant over-weights or under-weights the evidence, we use a parameter γ which intuitively captures how many "effective" red (n_r) or blue (n_b) chips the participant incorporates into their calculations, as in Equation 4. Thus, when

$\gamma = 1$, participants are neither over-weighting nor under-weighting the evidence; $\gamma < 1$ indicates conservatism while $\gamma > 1$ indicates over-weighting the data.³

$$P(n_r, n_b | x) \propto x^{\gamma n_r} (1-x)^{\gamma n_b} \quad (4)$$

As with the prior, the reported posteriors for each person were smoothed by adding 0.01 to the zero values and then normalising; this ensured that fits would not be artificially distorted by the propagation of zero values. (Analyses without this smoothing had qualitatively similar outcomes.) Optimal values for β and γ were calculated for the MAIN condition (the only one with both priors and posteriors) in aggregate as well as separately for each individual. The analysis was performed in R using the `optim` function, with β constrained to be within 0.0000001 and 0.9999999 and γ within 0.0000001 and 50 using the L-BFGS-B method. The function being minimised was the mean squared error between the model's prediction and the reported posterior. The supplemental materials contain information about the model fits, which were very good in all experiments.

Results

Aggregate performance

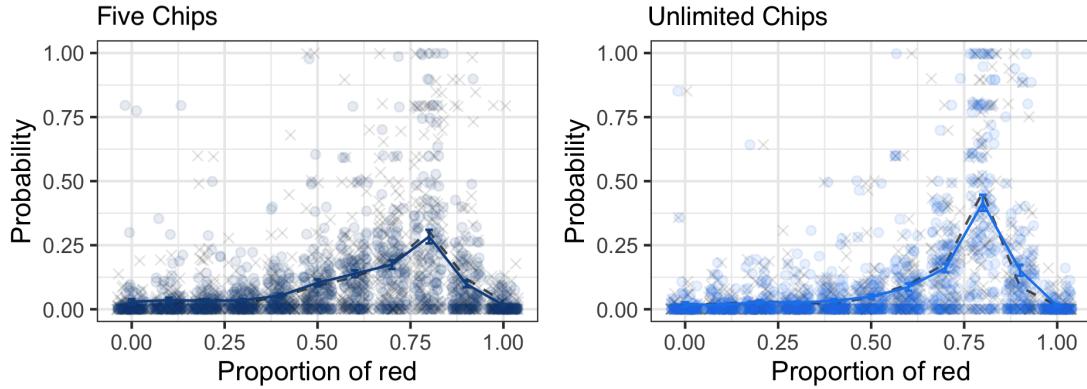
In order to ensure that the act of eliciting a prior or multiple posteriors did not change how participants reported probability distributions, we first compare the posteriors obtained from the two control conditions (i.e., the ONLYFIVE and ONLYUNLIMITED conditions) to the corresponding posteriors obtained from the MAIN condition (i.e. the posterior obtained after participants saw five chips and the posterior obtained after participants saw as many additional chips as they desired). As shown in Figure 3, the aggregate posteriors are extremely similar regardless of whether participants were asked to report their priors first (solid lines) or not (dotted lines). In both subplots, for all three conditions the mode is at 80%, indicating that participants on average correctly reported that they expected about 80% of the chips to be red regardless of the condition. Comparing the right subplot to the left subplot, we see that the peak was narrower indicating that the participants were more certain of the proportion of red chips after seeing more chips. Overall, this indicates that participants understood the task and reported reasonable distributions. More importantly, these results demonstrate that asking participants to estimate the prior did not substantially alter their subsequent estimates of the posterior.⁴ This allows us to focus on the MAIN condition, where each participant estimated three probability distributions: one before viewing any chips, one after viewing five chips, and one after viewing an unlimited number of additional chips.

³We also performed a version of this analysis which had two free parameters, one that weighted n_r and a separate one that weighted n_b , to capture a situation where participants might weight blue chips more or less than red chips. Results were qualitatively identical and the two parameters were similar to each other, suggesting that both chips were better modelled with one parameter.

⁴We also compared the average number of chips people asked for in the ONLYUNLIMITED condition ($M = 18.05$) and in the MAIN condition after viewing an unlimited number of chips ($M = 21.18$). A Welch t-test was not significant, $t(233.1) = 1.65, p = .100$, suggesting that participants who reported multiple distributions requested a similar number of chips as those who reported only one.

Figure 3

Experiment 1: Aggregate posterior estimates after viewing five or an unlimited number of chips



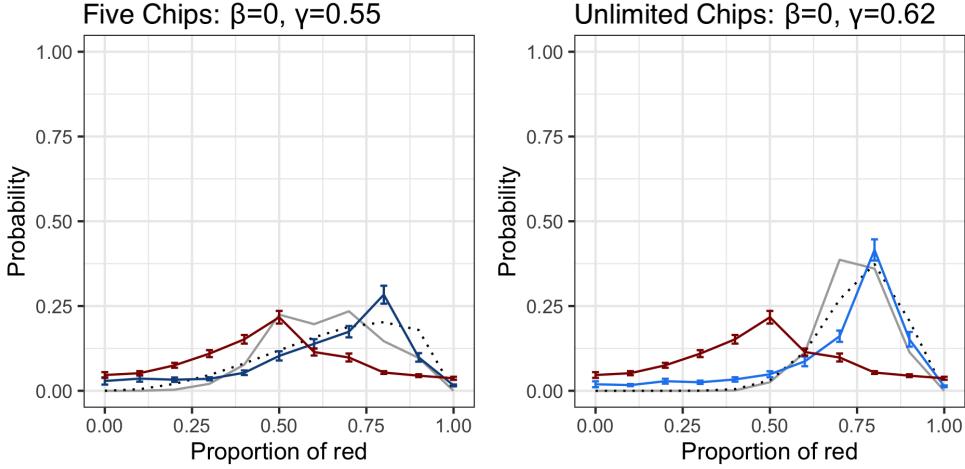
Note. Aggregate posterior estimates of the distribution of the probability that the bag contains a given proportion of red chips (y axis), for proportions ranging from 0% to 100% (x axis). *Left panel.* Posterior estimates after viewing five chips. The solid dark blue line reflects the aggregate posterior estimate of people in the MAIN condition after having seen five chips, while the dashed black line reflects the aggregate posterior estimate of those in the ONLYFIVE condition. Dark blue dots indicate individual estimates in the MAIN condition and light grey Xs indicate those in the ONLYFIVE condition. The aggregate posterior estimates are extremely similar in both conditions, with a mode at 80%, indicating that the posteriors are reasonable and that eliciting priors beforehand does not measurably change their behaviour. *Right panel.* Posterior estimates after viewing an unlimited number of chips. The solid light blue line reflects the aggregate posterior estimate of people in the MAIN condition after having seen unlimitted chips, while the dashed black line reflects the aggregate posterior estimate of those in the ONLYUNLIMITED condition. Light blue dots indicate individual estimates in the MAIN condition and light grey Xs indicate those in the ONLYUNLIMITED condition. The aggregate posterior estimates are extremely similar in both conditions, with a mode at 80%. As before, this indicates that the posteriors are reasonable and that reporting their distributions multiple times does not change their behaviour.

We can ask several questions of this data on the aggregate level. First, what prior distribution was reported? Participants were not given any information about the quantity of red or blue chips in the bag, so this question allows us to investigate what they presumed in the absence of any instruction. Figure 4 shows the aggregate prior (red line), which has a small peak at 50%, suggesting that on balance people think that a 50/50 split of red and blue chips is more likely than any other mixture. That said, the probability distribution is also fairly uniform across all possible values, indicating that participants would not be terribly surprised if the bag contained all red chips, all blue chips, or any of the other possible combinations.

A second question we can ask of the aggregate data is, when we fit it to our model by adjusting β and γ , what do the resulting parameters tell us about the degree of base rate neglect and conservatism shown by the population as a whole? As Figure 4 makes clear, the best-fit parameters

Figure 4

Experiment 1: Aggregate best-fit estimates in the MAIN condition after viewing five and an unlimited number of chips



Note. Aggregate distributions for prior beliefs (red line) as well as after seeing five chips (dark blue line, left panel) and unlimited chips (light blue line, right panel) in the MAIN condition. In both panels, the grey line indicates the optimal Bayesian prediction (i.e. $\beta = \gamma = 1$) given the aggregate prior, while the black dotted line indicates the prediction of the line of best fit based on the inferred parameters β and γ . In both panels, the best fit β is zero, indicating that the aggregate posterior was best fit by completely disregarding the aggregate prior (i.e., complete base rate neglect). The best fit values for γ indicates a moderate degree of conservatism in both conditions.

after both five chips and unlimited chips were similar. In both cases, they reflect that the aggregate posteriors were best captured assuming people ignore their reported priors completely (i.e., $\beta = 0$) and show a moderate degree of conservatism in updating (i.e., $\gamma < 1$). We can understand intuitively why this is the case by comparing the reported posteriors with the predicted posteriors that would we would expect from an optimal Bayesian reasoner (grey line, $\beta = \gamma = 1$). Such a reasoner would have a more bimodal posterior than we observed, especially after only five chips, which reflects the influence of the prior.

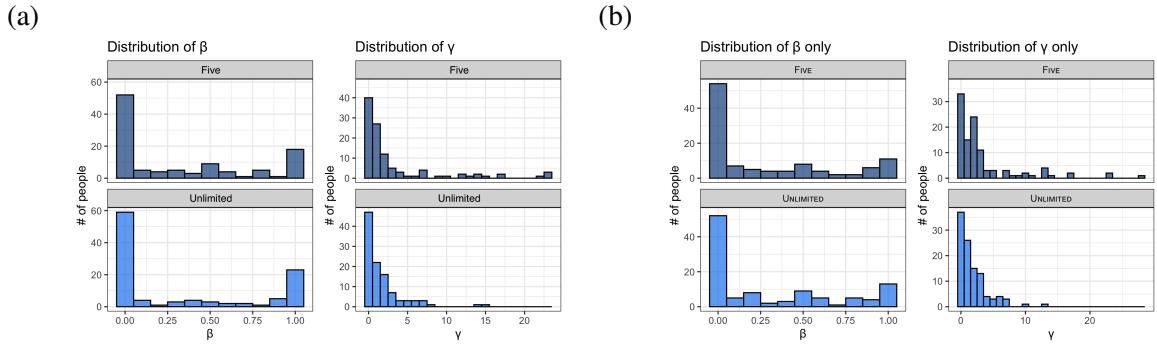
Individual performance

One of our main motivations was to understand how individuals (rather than populations) represented and updated their beliefs. Figure 5(a) shows the distribution of β and γ obtained when fit to each person separately. It is apparent that there is substantial individual variation and there are few differences based on whether five or unlimited chips were seen. That said, most people showed partial or complete base rate neglect: around half of people completely disregarded their priors (51.4% of people after seeing five chips and 56.1% after seeing unlimited chips had $\beta < 0.1$) and only a minority showed no base rate neglect at all (17.8% of people after seeing five chips and

22.4% after seeing unlimited chips had $\beta > 0.9$). Participants varied more in how they weighted the likelihood, with around half of the participants being conservative (50.4% of people after seeing five chips and 57.9% after seeing unlimited chips had $\gamma < 1$).

Figure 5

Experiment 1: Distribution of fitted β and γ values across individuals in the MAIN condition



Note. Histograms showing the distribution of best-fit β and γ values across individuals after five and unlimited chips in the MAIN condition. (a) β and γ were fit jointly; (b) β and γ were fit assuming the other was 1. Regardless of whether the parameters were fit jointly or not, distributions of β indicate that the majority of people showed a moderate or large amount of base rate neglect; their inferences were best described with β values less than one and often close to zero, indicating that the posterior distribution they reported was best explained by assuming that they disregarded their reported prior at least partially and often completely. There was a more varied distribution of γ values, with about half showing conservative updating ($\gamma < 1$).

There was no obvious systematic relationship between β and γ values within individuals; it was not the case that a low β meant high γ or vice versa (Spearman correlation, after five chips: $\rho = .040, p = .680$; after unlimited chips: $\rho = .18, p = .060$; see the supplemental materials for the scatterplots). In order to ensure that β and γ were identifiable by the model, we also separately fit each parameter while holding the other constant. That is, we (i) examined the range of β values obtained when the model assumed that each person updated their likelihood appropriately (i.e., that $\gamma = 1$ for everybody); and (ii) also examined the range of γ values obtained when the model assumed that each person updated their prior appropriately (i.e., that $\beta = 1$ for everybody). This result, shown in Figure 5(b), demonstrates that this makes no qualitative difference: the majority of participants still showed base rate neglect (i.e., were best fit by β values close to zero: 55.1% after five and unlimited chips had $\beta < 0.1$) and similar levels of conservatism (38.3% after five chips and 49.5% after unlimited chips had $\gamma < 1$).

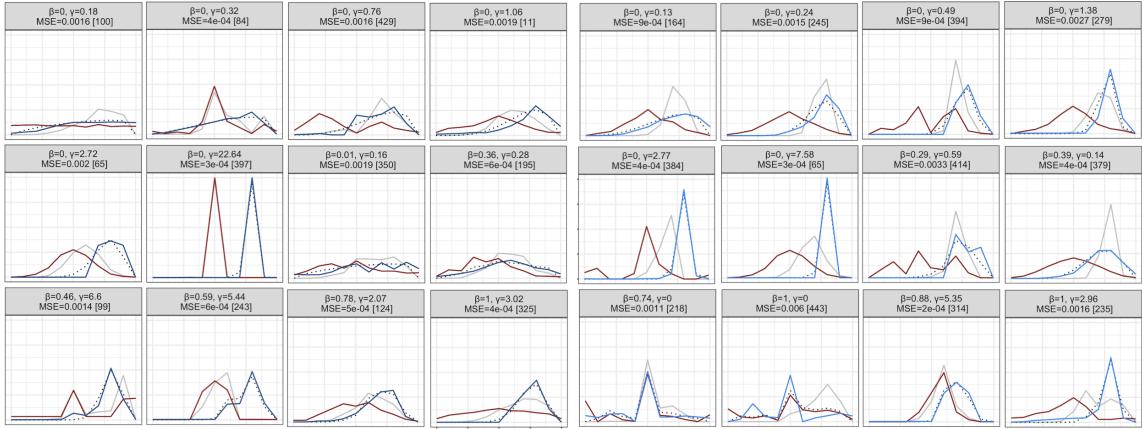
To get an intuitive sense of what people are doing, we can inspect the individual distributions. Figure 6 shows some representative examples, and all participants are shown in the supplemental materials. There is considerable heterogeneity: people report a wide variety of both priors and posteriors. That said, observation of the distributions makes it clear how it is possible to tease apart the weightings of the prior and the likelihood separately. Under-weighting the prior results

Figure 6

Experiment 1: Illustrative examples of individual distributions, data from the MAIN condition

Example individuals: Five chips

Example individuals: Unlimited chips



Note. A sample of some individual participants after receiving five chips (left) and unlimited chips (right). The red line is the reported prior, the dark and light blue lines are the reported posteriors after five and unlimited chips respectively, the grey line is the posterior obtained by a optimally calibrated Bayesian reasoner with that prior ($\beta = \gamma = 1$), and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel reports those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in parenthesis is the participant ID.

in a posterior distribution with a different shape (with multiple modes) or a different mode (closer to the likelihood) than the posterior distribution produced by an optimal Bayesian learner with that prior. By contrast, different likelihood weights change the height of the mode: conservative updating results in a mode that is lower than the Bayesian prediction, while over-weighting the likelihood results in a mode that is higher than the predicted one. As such, inspection of the individual curves is useful for understanding qualitatively what the quantitative fits of β and γ reveal.

Although our model fits in general were excellent (82.2% of people were fit with an MSE of 0.01 or less and 96.7% with 0.05 or less), one might still worry about whether our results were driven in part by the participants who were not fit well by the model. For instance, if all of the people for whom $\beta = 0$ were also fit badly, this might not mean that most people showed base rate neglect after all. In order to ensure that this was not the case, we redid all analyses after excluding the people with mean squared error greater than than 0.01. This did not change the qualitative results, with most of the 76 remaining participants still showing a high degree of base rate neglect and varying in conservatism (see the supplemental materials). This suggests that our results are not an artefact of poor fits, and we can be somewhat confident in our interpretation of the parameters.

Experiment 2

In Experiment 1 most participants showed base rate neglect, partially or completely ignoring their own reported prior when updating their beliefs in light of new data. Why did they do this? One possibility is that the task demands encouraged them to do so, since no prior was ever explicitly given and physically seeing chips being drawn may have made the data more salient. In Experiment 2, we investigated this possibility by explicitly giving participants the prior. People in the PEAKED prior condition were shown a distribution with a mode at a proportion of 80% red chips (as this most closely aligned with the data the participants would subsequently receive). Those in the UNIFORM prior condition were shown a completely flat prior; this is a useful comparison because reasoning based on this prior is equivalent to reasoning that completely ignores the prior. As a result, if participants always ignore their prior then the posteriors they report should be the same in both conditions; if not, the posterior should be sharper in the PEAKED condition.

Method

Participants

300 people (184 male, 113 female, 3 non-binary) were recruited via Prolific Academic and paid 60 British pence. Mean age was 26 years old. Sixty-one people were excluded because they either failed the bot check or did not adjust any bars when estimating distributions.

Materials and Procedure

This experiment involved the same procedure and instructions as Experiment 1 except that we presented participants with an explicit prior distribution using the same “bar” format that they used to report their own. In the PEAKED condition ($N = 121$) people were informed that a previous participant who had completed the task several times had stated that “There were usually about four times more red chips than blue chips in the bag (like, 80% red)” and had also drawn the plot in the left panel of Figure 7 to illustrate their statement. Conversely, the people in the UNIFORM condition ($N = 118$) were informed that a previous participant stated that “The number of red and blue chips in the bag keeps changing, doesn’t seem to be a pattern to it” and had drawn the plot in the right panel of Figure 7 to illustrate their statement.

Because Experiment 2 presented participants with an explicit prior, the procedure did not involve a prior elicitation step. Instead, after having been told the prior, participants were shown five chips (four red and one blue in random order, as before) and were asked to draw their posterior.

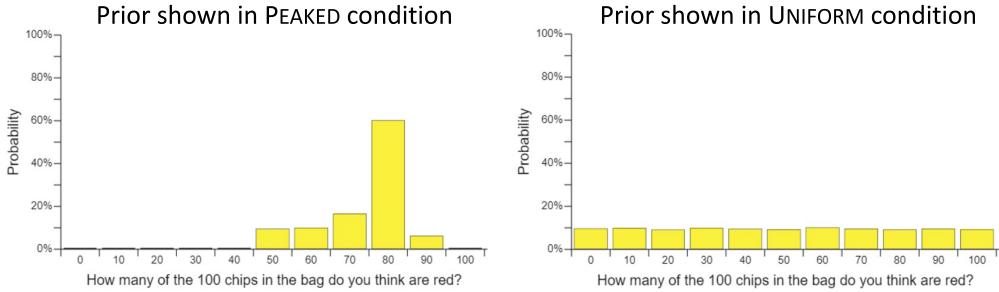
Results

Aggregate performance

We first present the aggregate distributions in each condition, along with the best-fit β and γ values. As Figure 8 shows, participants in the PEAKED condition were not entirely ignoring the

Figure 7

Priors shown to people in Experiment 2



Note. A screenshot depicting the priors that participants saw in the two conditions of Experiment 2.

prior; their posterior is tighter and sharper than in the UNIFORM condition, as one would expect if they were taking the prior into account. That said, a comparison to the posterior inferred by an optimal Bayesian – along with the inferred β and γ values – demonstrates that people still showed substantial base rate neglect (albeit less than before) and some degree of conservatism.

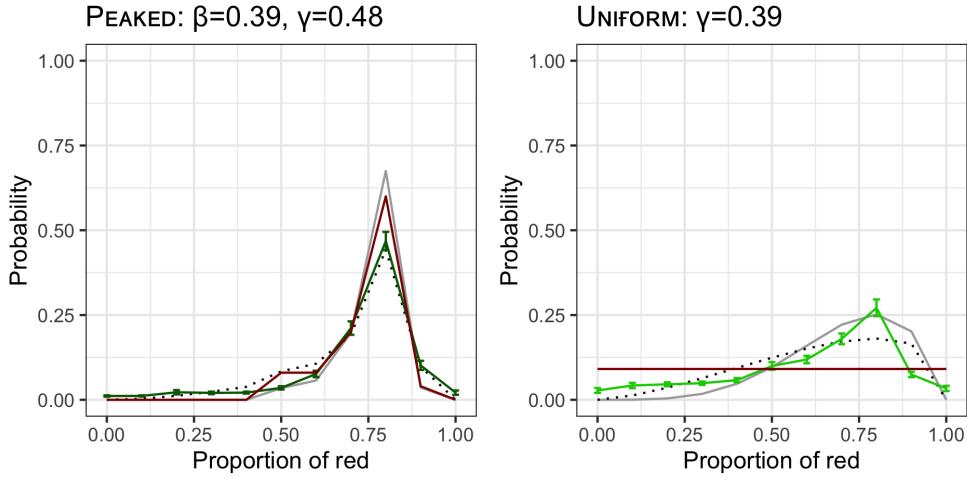
Individual performance

As before, we performed individual-level analyses by fitting each participant to the value of β and γ that best captured their reported posterior based on the prior they were given. The distribution of these parameters in each condition is shown in Figure 9(a) (recall that there are no β values in the UNIFORM condition because the given prior was uniform). There is again substantial individual variation, but most people in the PEAKED condition showed partial or complete base rate neglect: 38.8% of participants disregarded their priors (with $\beta < 0.1$) and only 3.3% showed no base rate neglect at all (with $\beta > 0.9$). That said, more participants than in Experiment 1 paid *some* attention to the prior they were given, even if they did not weight it as strongly as an optimal Bayesian reasoner would have. The degree to which participants weighted the likelihood depended on their condition. Participants in the PEAKED condition were less likely to be conservative than those in the UNIFORM condition: 38.8% in PEAKED and 61.0% in UNIFORM had $\gamma < 1$.

However, we must sound an important cautionary note at this point. The distributions in Figure 8 indicate that, unlike in Experiment 1, it may not be so easy to disentangle the effects of β and γ in the PEAKED condition of Experiment 2. That is because increasing the weighting of the prior and increasing the weighting of the likelihood both have the same effect: to sharpen the posterior at 80%. This suggests that in the PEAKED condition, β and γ may not be uniquely identifiable: participants with peaked posteriors might be equally well-fit by a high β and low γ , low γ and high β , or moderate γ and moderate β . (Note that this concern does not apply to the UNIFORM condition as β did not need to be estimated in that condition.) As in Experiment 1, we did not find an obvious systematic relationship between β and γ values within individuals, but the identifiability issue means that it is difficult to interpret this result (Spearman correlation, $\rho = .141, p = .124$; see

Figure 8

Experiment 2: Aggregate best-fit estimates



Note. Reported distributions for prior beliefs (red line) and posteriors in the PEAKED (dark green line, left panel) and UNIFORM (light green line, right panel) conditions. The grey line indicates the optimal Bayesian prediction given the prior, while the black dotted line indicates the prediction of the line of best fit based on the inferred parameters β and γ . The posterior distribution is noticeably sharper in the PEAKED condition, suggesting that people are using the base rate information to at least some extent. Consistent with this, β in the PEAKED condition is 0.39, indicating that the aggregate posterior was best fit assuming that participants partially used the prior they were given: more than in Experiment 1, but less than an optimal Bayesian would ($\beta = 1$). In both conditions, the value for γ indicates a moderate degree of conservatism on average.

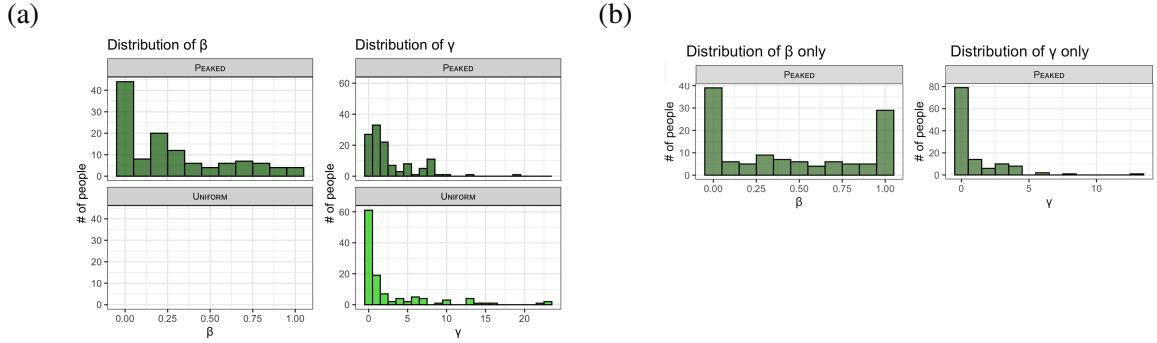
the supplemental materials for the scatterplots and further discussion).

To explore the extent of the identifiability problem, we again (i) examined the β values obtained when the model assumed that each person updated their likelihood appropriately (i.e., that $\gamma = 1$ for everybody); and (ii) examined the γ values obtained when the model assumed that each person updated their prior appropriately (i.e., that $\beta = 1$ for everybody). (Note that this was only relevant for the PEAKED condition, since in the uniform condition there was only one parameter anyway). The result, shown in Figure 9(b), demonstrates that the parameters are distributed differently when inferred singly rather than jointly. When γ is constrained to 1, more participants showed no base rate neglect at all (25.6% instead of 3.3% obtained $\beta > 0.9$) and slightly fewer showed complete base rate neglect (33.9% instead of 38.8% had $\beta < 0.1$). Conversely, when β is constrained to be 1, more participants appear to be reasoning conservatively (74.4% instead of 38.8% had $\gamma < 1$). This suggests that for participants who made reasonably strong inferences, the model could not determine whether this was because they weighted their prior highly and reasoned conservatively, they ignored their prior and weighted the likelihood highly, or they weighted both a moderate amount.

Although our model fits were again excellent (79.1% of people were fit with MSE less than 0.01, and 98.7% with MSE less than 0.05), as before we redid all analyses after excluding the

Figure 9

Experiment 2: Distribution of fitted β and γ values across individuals



Note. Histograms showing the distribution of best-fit β and γ values across individuals in the PEAKED and UNIFORM conditions. (a) β and γ were fit jointly; (b) β and γ were fit assuming the other was 1. When the parameters were fit jointly (a), most people (although fewer than in Experiment 1) showed a moderate or large amount of base rate neglect and there was a varied distribution of γ values, with more people showing conservative updating in the UNIFORM condition. When they were not (b), the parameters look far different: many more people in the PEAKED condition show no base rate neglect (β at or near 1) and many more are conservative ($\gamma < 1$). This suggests that in Experiment 2, γ and β are not easily distinguishable by the model. (N.B. This analysis was done only for the PEAKED condition as β was not estimated in the UNIFORM condition.)

people that were not fit well by our model (i.e., the people with mean squared error greater than than 0.01). As documented in the supplemental materials, this did not change the qualitative results: the remaining 189 people still appeared to show some base rate neglect in the aggregate, but the PEAKED condition had a sharper posterior than the UNIFORM condition, demonstrating that they did take the prior into account at least somewhat. For individual participants, the best-fit β and γ values were, again, not identifiable in the PEAKED condition unless the inferences were very weak (in which both β and γ were very low).

Experiment 3

The PEAKED condition of Experiment 2 suggested that even when the prior is made explicit, people underweight it relative to how they should weight it according to Bayes' theorem. However, this finding must be interpreted with caution because, in that condition, we were not able to reliably disentangle the effects of weighting the prior from the effects of weighting the likelihood since both peaked at a proportion of 80% red. Experiment 3 addressed this concern by explicitly giving participants a prior that peaked at a value of 50%.

Method

Participants

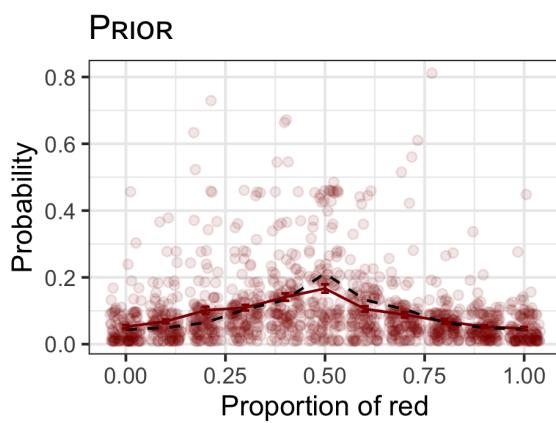
300 participants (132 male, 162 female, 6 non-binary) were recruited via Prolific Academic and paid 60 British pence. Mean age was 25 years old. Thirty-nine people were excluded because they either failed the bot check or did not adjust any bars when estimating distributions.

Materials and Procedure

The experiment involved the same procedure and instructions as before except for the following differences. In the ESTIMATED condition participants were not provided with a prior. It was thus identical to the MAIN condition of Experiment 1, except that the experiment stopped after the participants had reported the first posterior (i.e. after the participant had seen five chips). In the GIVEN condition people ($N = 133$) were provided with a prior. It was thus identical to the PEAKED condition of Experiment 2 except that the prior they were shown corresponded to the aggregate prior reported in Experiment 1. We designed it this way because it means that in both conditions we would expect people to have the same prior; the conditions differ only in whether that prior was explicitly provided or not.

Figure 10

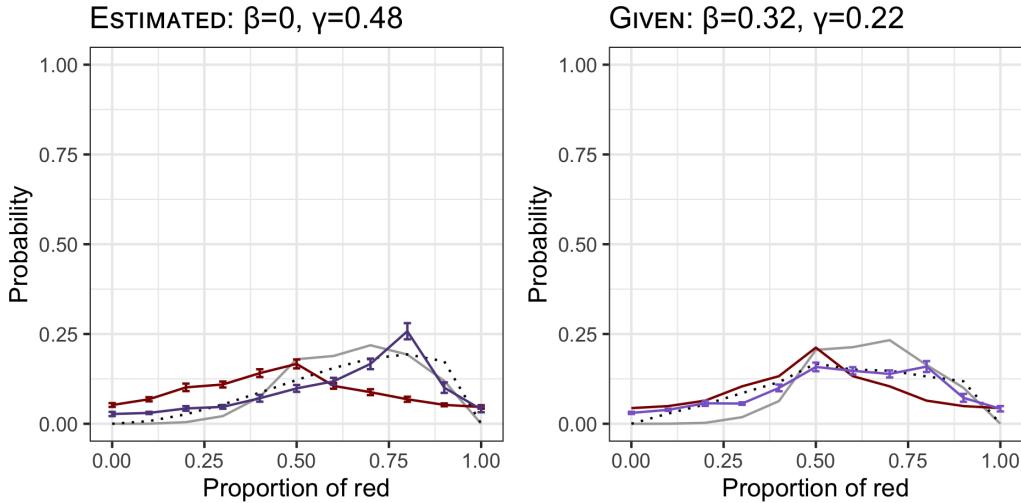
Experiment 3: Reported prior distributions



Note. Aggregate prior estimates of the distribution of the probability that the bag contains a given proportion of red chips (y axis), for proportions ranging from 0% to 100% (x axis). The solid red line reflects the aggregate of the priors reported in the ESTIMATED condition, while for comparison the dashed black line reflects the prior shown to people in the GIVEN condition. Red dots indicate individual estimates in the ESTIMATED condition. The priors are extremely similar in both conditions, suggesting that any difference in posteriors is not due to differences in the prior.

Figure 11

Experiment 3: Aggregate best-fit estimates



Note. Reported distributions for prior beliefs (red line) and posteriors in the ESTIMATED (dark purple line, left panel) and GIVEN (light purple line, right panel) conditions. The grey line indicates the optimal Bayesian prediction given the prior, while the black dotted line indicates the prediction of the line of best fit based on the inferred parameters β and γ . The posterior distribution is unimodal in the ESTIMATED condition but multimodal in the GIVEN condition, suggesting that in the GIVEN condition people are using the given prior, at least some extent. Consistent with this, β in the GIVEN condition is 0.32, indicating that the aggregate posterior was best fit assuming that participants partially used the prior they were given: more than in Experiment 1 and the ESTIMATED condition where β was always equal to zero, but less than an optimal Bayesian would ($\beta = 1$). In both conditions, the value for γ indicates a moderate degree of conservatism, though the degree of conservatism is somewhat less in the ESTIMATED condition.

Results

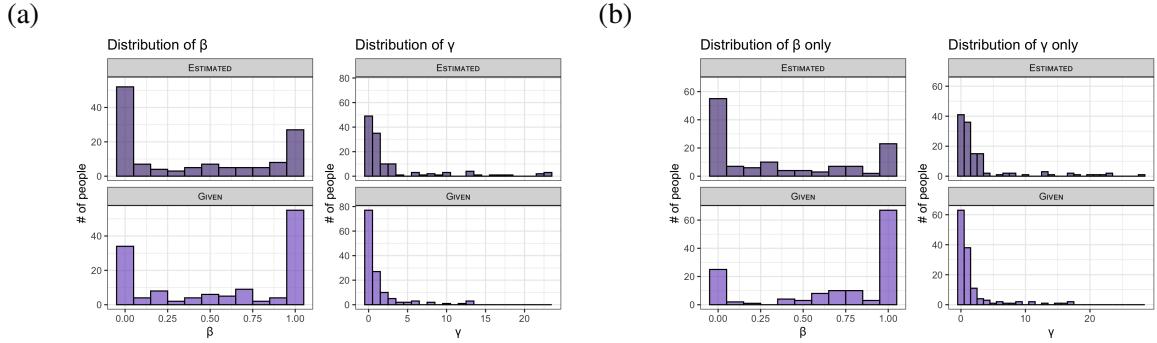
Aggregate performance

As shown by Figure 10, the prior reported by the participants in the ESTIMATED condition (solid red line) was very similar to the prior provided to the participants in the GIVEN condition (dashed black line). This suggests that any differences in the posteriors in the two conditions is unlikely to be due to differences in their priors.

Figure 11 shows the aggregate posterior distributions for each condition, shown alongside the optimal Bayesian prediction as well as the prediction made using the best-fit parameters β and γ . As expected, the best fit parameters for the ESTIMATED condition ($\beta = 0, \gamma = 0.48$) are very similar to the best fit parameters in the FIVE condition in Experiment 1 ($\beta = 0, \gamma = 0.55$), with participants in the aggregate demonstrating complete base rate neglect ($\beta = 0$). Conversely, in the GIVEN condition participants made much more use of the prior ($\beta = 0.32$). This resulted in a posterior with two modes corresponding to the peaks of the prior and the likelihood. This is

Figure 12

Experiment 3: Distribution of fitted β and γ values across individuals



Note. Histograms showing the distribution of best-fit β and γ values across individuals in the ESTIMATED and GIVEN conditions. (a) β and γ were fit jointly; (b) β and γ were fit assuming the other was 1. In both cases, the distributions for β show that most people in the ESTIMATED condition, as in Experiment 1, showed a moderate or large amount of base rate neglect, but that this flipped in the GIVEN condition. There was again a varied distribution of γ values.

consistent with the finding from the PEAKED condition of Experiment 2 that when the prior is made explicit, participants make use of it, but not to the extent predicted by Bayes' theorem.

Individual performance

As before, we performed individual-level analyses by fitting each participant to the value of β and γ that best captured their reported posterior given their prior. The distribution of these parameters in each condition is shown in Figure 12(a). The results in the ESTIMATED condition are very similar to the analogous condition of Experiment 1, with many participants disregarding their prior (43.8% had a $\beta < 0.1$, compared to 51.4% previously) and a minority weighting it appropriately (23.4% had a $\beta > 0.9$, compared to 17.8% previously). The results from the GIVEN condition are consistent with the observation from Experiment 2 that participants pay more attention to the base rate when the prior is made explicit: fewer people in the GIVEN condition than the ESTIMATED one ignored the base rate entirely (26.3% had $\beta < 0.1$) and more weighted it appropriately (41.4% had $\beta > 0.9$). As before, a moderate number of participants reasoned conservatively (51.6% in the ESTIMATED condition and 70.7% in the GIVEN condition had $\gamma < 1$). There was also again no obvious systematic relationship between β and γ (Spearman correlation, ESTIMATED: $\rho = .137, p = .124$; GIVEN: $\rho = -.04, p = .675$; see supplemental materials for scatterplots). Thus, the degree to which an individual weights the prior does not predict the degree to which they weight the likelihood.

To confirm that we could now disentangle the effects of β and γ , we again examined (i) the range of β values obtained when the model assumed that each person updated their likelihood appropriately (i.e., that $\gamma = 1$ for everybody) and, in a separate analysis, (ii) examined the range

of γ values obtained when the model assumed that each person updated their prior appropriately (i.e., that $\beta = 1$ for everybody). The results are shown in Figure 12(b). As in Experiment 1, the distributions obtained by fitting the parameters jointly or singularly are similar. In the ESTIMATED condition, a large number of participants show complete base rate neglect (46.1% have $\beta < 0.1$) and few weight it appropriately (18.8% have $\beta > 0.9$). Conversely, fewer participants in the GIVEN condition show complete base rate neglect (20.3% have $\beta < 0.1$) and more weight it appropriately (51.1% have $\beta > 0.9$). Qualitatively similar numbers of participants are conservative as when the parameters are inferred jointly (63.9% in GIVEN and 43% in ESTIMATED had $\gamma < 1$).

The model fits for Experiment 3 were equally as good as in previous experiments (81.6% of people had an MSE of less than 0.01, and 98.5% less than 0.05). Nevertheless, as before, we redid all analyses after excluding the people with MSE less than 0.01, leaving 213 in the dataset. As shown in the supplementary materials, this did not change the qualitative results. On the aggregate as well as individual levels, in the ESTIMATED condition participants were more likely to ignore their prior whereas in the GIVEN condition more participants used the prior.

Discussion

In this paper we asked to what extent human probability reasoning conforms to the normative standards prescribed by Bayes' theorem when participants present their probability estimates as entire distributions rather than as point estimates. Our first experiment was inspired by the standard balls-and-urn task. Participants were shown a bag containing a number of chips, some red and some blue, and were asked to provide three probability distributions (one prior and two posteriors) using a visual histogram tool similar to that of Goldstein and Rothschild (2014). The task description gave no information about the likely ratio of red to blue chips, and participants varied substantially in the shape of the priors they assumed, though many were peaked, to a greater or lesser extent, around 50%. Fitting individual participants revealed that, regardless of whether they saw only five chips or were allowed to view as many chips as they desired, the majority showed substantial base rate neglect (i.e., ignoring the prior they had reported) and varied in the degree to which they were conservative (i.e., updating their likelihoods less than a normative Bayesian reasoner).

In order to determine whether people ignored their prior because it was not explicitly stated, in Experiment 2 we presented people with either a uniform or a peaked prior and then asked for their posterior distributions after seeing five chips. Here the aggregate results revealed that, even when given an explicit prior, there was some degree of base rate neglect. However, they had sharper posteriors when given a peaked prior than when given a uniform prior, indicating that the priors were not being ignored entirely. This was supported by fitting individual participants: although variation was again substantial, more people used the prior when it was made explicit in the PEAKED condition than when it was not.

This final result needs to be interpreted with caution as it was unclear to what extent the two parameters could be uniquely identified for each individual in the PEAKED condition: since both the prior and the likelihood peaked at the same proportion (80%), the effect of one could easily be

mistaken as being caused by the other. Experiment 3 addressed this concern by using an explicit prior that peaked at a different proportion, thereby ensuring that the effects of the prior and the likelihood were uniquely identifiable. It confirmed the findings of Experiment 2: when the prior is made explicit, people weigh it more than when it is not.

Some of the subtleties that arose in our analysis illustrate both the benefits and complexities of measuring and fitting full probability distributions. There are several benefits. For instance, providing the prior and the likelihood have different modes, this method allows us to disentangle, for a single individual participant on a single reasoning problem, to what extent they under-weight or over-weight both their prior *and* their likelihood. This is not possible using any other methodology: mathematically, a single point posterior is compatible with any possible weighting of each. The studies that do attempt to disentangle prior and likelihood weightings do so by presenting multiple problems, systematically varying both the priors and the evidence (e.g., Benjamin et al., 2019; Griffin & Tversky, 1992). This is sometimes useful, but presumes that people weigh their priors and likelihoods similarly across all problems. Our results suggest that this is not necessarily the case: people showed more base rate neglect in some circumstances than in others.

Another benefit is that because each individual was fit separately for both prior and likelihood weights, we could determine both how each of these tendencies varied among people as well as how they related to each other. For instance, Experiment 1 demonstrated that most people either completely ignored their priors (with β close to 0) or weighted them appropriately (with β close to 1); that is, the distribution over β was bimodal, with few intermediate values. Because of the identifiability issue we cannot draw strong conclusions from the PEAKED condition of Experiment 2, but it is striking then when the likelihood weight was held constant, a similar bimodal distribution emerged. This finding was then confirmed by Experiment 3, which did not suffer from the identifiability issue. Moreover, across all experiments we observed no systematic relationship between how people weighted their priors and likelihood terms: it wasn't the case that people who underweighted one tended to under-weight the other, or vice-versa. To our knowledge this is a novel finding; future research is necessary to determine how robust it is and how far it extends.

More broadly, this research demonstrates *why* it can be useful to elicit and analyse entire distributions rather than single point estimates. As long as (i) the prior is not uniform and (ii) the prior and likelihood have different modes (unlike in Experiment 2), the two terms make different contributions to the shape of the posterior distribution, so their individual contributions can be estimated. Our approach can be extended to consider cases like Figure 1 where the differences are even more pronounced than in the present study. There is a great deal of potential in applying this methodology to long-standing problems in human reasoning. Might the framing of the problem (i.e. how the problem is presented to the participants) affect base rate neglect (Barbey & Sloman, 2007) at least in part because people may implicitly assume priors with different distributional shapes (reflecting different levels of confidence or extent) depending on how the problem is presented to them? Might base rate neglect be smaller for priors that are easier to use, represent, or sample from? To what extent do anchoring effects change if the information is presented as a full distribution? Do

the same individuals weight their priors and likelihoods the same across different problems? These are only some of the questions that can now be addressed.

One potential worry about the validity of our method is the extent to which people can actually accurately report their underlying distribution. If people reason by drawing a small number of samples from their distribution, as some suggest (Vul et al., 2014), it is not obvious that this would be sufficient for people to reconstruct and report the actual distribution. Although this is a possibility we cannot rule out with certainty, it seems unlikely: the distributions people reported seem reasonable both individually and in the aggregate, and reflect the overall patterns one would expect: tightening with additional information in Experiment 1, stronger inferences with a stronger prior in Experiment 2, and more reliance on the prior when it is made explicit in Experiment 3. Moreover, the method has been tested and shown to be equally if not more accurate than verbal reports (Goldstein & Rothschild, 2014). That said, it remains to be seen if it can be applied to more complex distributions: for instance, bimodal or highly skewed distributions may require more than a few samples in order to capture the important moments, and it is unclear whether people can report or use such distributions while reasoning. This is another exciting avenue of future research that our methodology opens up.

In sum, this paper presents initial research demonstrating the utility of eliciting and fitting full distributions when studying probabilistic reasoning. Across three experiments, we found substantial variation in the extent to which people showed base rate neglect and conservatism, which our method allowed us to measure in individuals on single problems. While most people tended to disregard the base rate, they did so less when it was explicitly presented. Moreover, there was no apparent systematic relationship between base rate neglect and conservatism within individuals.

References

- Barbey, A., & Sloman, S. (2007). Base-rate neglect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30(3), 241–297. <https://doi.org/10.1017/S0140525X07001653>
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211–233. [https://doi.org/10.1016/0001-6918\(80\)90046-3](https://doi.org/10.1016/0001-6918(80)90046-3)
- Benjamin, D., Bodoh-Creed, A., & Rabin, M. (2019). Base-rate neglect: Foundations and implications. *Working Paper*.
- Corner, A., Harris, A., & Hahn, U. (2010). Conservatism in belief revision and participant skepticism. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 41625–1630). Cognitive Science Society.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction - frequency formats. *Psychological Review*, 102(4), 684–704.
- Goldstein, D., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, 9(1), 1–14.

- Grether, D. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics*, 95(3), 537–557. <https://doi.org/10.2307/1885092>
- Grether, D. (1992). Testing Bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization*, 17(1), 31–57. [https://doi.org/10.1016/0167-2681\(92\)90078-p](https://doi.org/10.1016/0167-2681(92)90078-p)
- Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24(3), 411–435. [https://doi.org/10.1016/0010-0285\(92\)90013-r](https://doi.org/10.1016/0010-0285(92)90013-r)
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14, 357–364.
- Hammerton, M. (1973). A case for radical probability estimation. *Journal of Experimental Psychology*, 101(2), 252–254.
- Holt, C., & Smith, A. (2009). An update on Bayesian updating. *Journal of Economic Behavior and Organization*, 69, 125–134.
- Johnson, N., & Kotz, S. (1977). *Urn models and their application: An approach to modern discrete probability theory*. Wiley.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251. <https://doi.org/10.1037/h0034747>
- Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, 119(4), 685–722.
- Kennedy, M., Willis, W., & Faust, D. (1997). The base-rate fallacy in school psychology. *Journal of Psychoeducational Assessment*, 15(4), 292–307.
- Lieder, F., & Griffiths, T. L. (2020). Advancing rational analysis to the algorithmic level. *Behavioral and Brain Sciences*, 43, e27. <https://doi.org/10.1017/S0140525X19002012>
- Lucas, C., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science*, 34, 113–147.
- Mandel, D. (2014). The psychology of Bayesian reasoning. *Frontiers in Psychology*, 5, 1144.
- Mozer, M., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, 32(7), 1133–1147. <https://doi.org/10.1080/03640210802353016>
- Peterson, C., & Miller, A. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, 70, 117–121. <https://doi.org/10.1037/h0022023>
- Phillips, L., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72(3), 346–354. <https://doi.org/10.1037/h0023653>
- Sanborn, A., Griffiths, T. L., & Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4), 1144–1167.
- Savage, L. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336), 783–801. <https://doi.org/10.2307/2284229>

- Schlag, Tremewan, J., & van der Weele, J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, 18(3), 457–490. <https://doi.org/10.2139/ssrn.2353295>
- Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 6(6), 649–744.
- Tribe, L. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, 84(6), 1329–1393.
- Vincent, B. (2015). A tutorial on Bayesian models of perception. *Journal of Mathematical Psychology*, 66, 103–114.
- Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4), 599–637. <https://doi.org/10.1111/cogs.12101>
- Vul, E., Hanus, D., & Kanwisher, N. (2009). Attention as inference: Selection is probabilistic; responses are all-or-none samples. *Psychological Science*, 138(4), 546–560. <https://doi.org/10.1037/a0017352>
- Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7), 645–647. <https://doi.org/10.1111/j.1467-9280.2008.02136.x>
- Wallsten, T., & Budescu, D. (1983). State of the art—encoding subjective probabilities: A psychological and psychometric review. *Management Science*, 29(2), 151–173. <https://doi.org/10.1287/mnsc.29.2.151>
- Wolpert, D. (2009). Probabilistic models in human sensorimotor control. *Human Movement Science*, 26(4), 511–524.