

Beliefs and Bayesian reasoning

Andrew L. Cohen¹ · Sara Sidlowski¹ · Adrian Staub¹

Published online: 7 September 2016
© Psychonomic Society, Inc. 2016

Abstract We examine whether judgments of posterior probabilities in Bayesian reasoning problems are affected by reasoners' beliefs about corresponding real-world probabilities. In an internet-based task, participants were asked to determine the probability that a hypothesis is true (*posterior probability*, e.g., a person has a disease, given a positive medical test) based on relevant probabilities (e.g., that any person has the disease and the true and false positive rates of the test). We varied whether the correct posterior probability was close to, or far from, independent intuitive estimates of the corresponding 'real-world' probability. Responses were substantially closer to the correct posterior when this value was close to the intuitive estimate. A model in which the response is a weighted sum of the intuitive estimate and an additive combination of the probabilities provides an excellent account of the results.

Keywords Bayesian reasoning · Belief bias · Modeling

Beliefs about the real world can affect reasoning. A well-known example is the belief bias in categorical syllogistic reasoning: a logical inference is more likely to be viewed as valid if the conclusion is believable in light of real-world knowledge (Evans, Barston, & Pollard, 1983). Here, we reach a similar conclusion regarding Bayesian reasoning, showing that this form of reasoning is influenced by a pre-experimental, intuitive estimate of the posterior probability.

In a typical Bayesian reasoning problem, participants are provided with three probabilities in the following form (adapted from Eddy, 1982; Gigerenzer & Hoffrage, 1995):

The probability that a person has breast cancer is 1 %. [*Base rate, $p(H)$*]

If a person has breast cancer, the probability that they will test positive is 85 %. [*True positive rate, $p(D|H)$*]

If a person does not have breast cancer, the probability that they will test positive is 15 %. [*False positive rate, $p(D|\neg H)$*]

The participant is asked to determine the *posterior* probability:

If a person tests positive, what is the probability they have breast cancer? [*Posterior, $p(H|D)$*]

Bayes' theorem provides the posterior probability as follows:

$$p(H|D) = \frac{p(D|H)p(H)}{p(D|H)p(H) + p(D|\neg H)p(\neg H)}$$

The correct posterior in this problem is 5.4 %. However, people typically perform very poorly on these problems (Bar-Hillel, 1980; Kahneman & Tversky, 1972). Even the responses of medical experts have been shown to be too high by about an order of magnitude on a similar problem (Eddy, 1982).

As in this example, real-world scenarios are often used as stimuli. For example, Gigerenzer and Hoffrage (1995) asked participants to estimate the probability that there is prenatal damage to a fetus given that the mother had German measles, and the probability that a child will develop bad posture given that he or she carries heavy books to school. The starting point

✉ Andrew L. Cohen
acohen@psych.umass.edu

¹ Department of Psychological and Brain Sciences, University of Massachusetts, 441 Tobin Hall, 135 Hicks Way, Amherst, MA 01003-7701, USA

of the present study is the observation that participants are likely to have pre-experimental beliefs about these posteriors, e.g., the conditional probability that a child will develop bad posture given that he or she carries heavy books. We call these pre-experimental beliefs *intuitive estimates*. It is plausible that Bayesian reasoning, like syllogistic reasoning, is influenced by how well the inferential conclusion accords with participants' beliefs about the real world (Anderson et al., 2015; Evans, Handley, Over, & Perham, 2002). We hypothesized that the posterior probabilities that participants provide reflect, in part, their intuitive estimates of the conditional probabilities that they are being asked to compute.

To address this question, we presented participants with Bayesian reasoning problems that shared common base, true positive, and false positive rates, but differed with respect to the real-world situation depicted in the problem. The actual posterior was either *believable*, i.e., relatively close to the mean intuitive estimate of the same conditional probability; *unbelievable*, i.e., far from the mean intuitive estimate; or *neutral*, i.e., subjects were not expected to have clear intuitive estimates.

Because intuitive estimates of the probability of breast cancer given a positive test result are typically very high (a mean of 62 % in the current experiment), the posterior in the problem above (5.4 %) is unbelievable. However, in the following example, with the same numerical values, the posterior is believable, as the mean intuitive estimate of the probability that a person who is coughing has pneumonia was only 13 %:

The probability that a person has pneumonia is 1 %
 If a person has pneumonia, the probability that they are coughing is 85 %
 If a person does not have pneumonia, the probability that they are coughing is 15 %

For this set, the neutral posterior question was “If an animal is a cat, what is the probability that it has tri-esterone?”. Our prediction was that estimated posteriors would be closer to the normatively correct Bayesian posteriors when the correct posterior was believable than when it was unbelievable.

Method

Participants

Fifty-four people participated through Amazon's Mechanical Turk (<https://www.mturk.com/mturk/>) and received US \$2.00. Assuming a two-tailed, dependent-measures *t*-test with $\alpha=0.05$ and $d=0.50$, 54 participants are needed to achieve a power of 0.95. Participants self-reported as fluent English speakers, in the United States, and between 18 and 74 years old.

Stimuli

Fifty sets of three Bayesian reasoning problems were initially generated. The probabilities within each set were identical. The base rates, true positive rates, and false positive rates for these problems, hereafter called the *rates*, were sampled randomly from uniform distributions in the ranges 0–59 %, 51–99 %, and 1–49 %, respectively. The correct posterior probabilities ranged from 5 % to 94 %. The content was selected so that, within each set, the correct posterior was judged by the experimenters to be believable in one case, unbelievable in one case, and neutral in one case. All rate statements were designed to be plausible.

A norming study with 108 participants (27 in each of four conditions), carried out on Mechanical Turk via Ibex Farm (<http://spellout.net/ibexfarm/>), served to select a subset of these 50 problem sets for use in the main experiment. The norming study ensured that the believability and plausibility assumptions were met for the selected items.

In the first condition of the norming study, participants viewed each of the 50 posterior prompts alone, e.g., “What is the probability that if a person is coughing they have pneumonia?”, and were asked to provide an estimate of this probability using radio buttons that ranged from 0 % to 100 % in 5 % increments, e.g., 11–15 %, 16–20 %, etc. The center of the response range (e.g., 13 % for the 11–15 % button) was used in statistical analysis. For each of the 50 sets, we calculated the mean absolute difference between the actual posterior and the intuitive estimate in the believable and unbelievable problems. The 24 sets with the largest differences between these values were selected for the main experiment. The mean distances to the actual posterior for the stimuli in each condition are provided in Table 1 (i.e., $|p_{ie}(H|D) - p(H|D)|$). The mean intuitive estimate is 0.29 closer to the correct posterior for the believable stimuli than for the unbelievable stimuli. The correlations of the intuitive estimate to the actual posterior for the selected stimuli in the believable, neutral, and unbelievable conditions were $r = 0.83, 0.10,$ and -0.8 , respectively.

In the remaining three norming conditions, participants viewed either the base, true positive, or false positive rate statements, and were asked to “rank the plausibility of this statement” on a 4-point scale (1 = very implausible, 2 = somewhat implausible, 3 = somewhat plausible, 4 = very plausible).¹ Fifty intuitively implausible rate statements were also included as fillers. The mean plausibility ratings for the rate statements in each condition of the 24 items that were selected for inclusion in the experiment are given in Table 1 (i.e., $p_{\text{plaus}}(H)$, $p_{\text{plaus}}(D|H)$, and $p_{\text{plaus}}(D|\neg H)$). The believable and unbelievable plausibility

¹ Full instructions for plausibility ratings: imagine that a friend tells you an interesting fact they just learned. This fact is the probability of a particular outcome or event. For example, ‘if someone is sneezing the probability that they have a cold is 55 %’. Your job is to figure out how plausible the fact is.

Table 1 Mean plausibility ratings for each rate statement and mean absolute distance of intuitive estimates and estimated posteriors to the actual posterior for the stimuli in each believability condition

Condition	$p_{\text{plaus}}(H)$	$p_{\text{plaus}}(D H)$	$p_{\text{plaus}}(D \neg H)$	$ p_{\text{ie}}(H D) - p(H D) $	$ p_{\text{est}}(H D) - p(H D) $
Believable	2.94 (0.26)	3.19 (0.40)	2.57 (0.30)	0.21 (0.06)	0.24 (0.20)
Neutral	2.65 (0.36)	2.76 (0.35)	2.35 (0.27)	0.32 (0.12)	0.25 (0.21)
Unbelievable	2.80 (0.35)	3.10 (0.41)	2.28 (0.37)	0.50 (0.13)	0.32 (0.23)

p_{plaus} Plausibility rating during the norming study, $p_{\text{ie}}(H|D)$ intuitive estimate of the posterior during the norming study, $p_{\text{est}}(H|D)$ estimate of the posterior in the main experiment, $p(H|D)$ actual posterior. Plausibility estimates could range from 1 to 4 (1 = very implausible; 4 = very plausible). The differences could range from 0 to 1. Values in parentheses are standard deviations

ratings for the base rate, $t(23) = 1.45$, $p = .16$, and the true positive rate, $t(23) = 0.85$, $p = .40$, were not statistically different. The plausibility of the false positive rate was somewhat lower in the unbelievable condition, $t(23) = 3.16$, $p = .004$. We note, however, that the difference was small in absolute terms, and that the false positive rate in the unbelievable condition did not differ significantly from the neutral condition, $t(23) = 0.71$, $p = .49$.

Three experimental lists were created from the 24 stimulus sets, so that each participant in the main experiment saw one problem from each of the 24 sets, with eight in each believability condition. Lists were created by first ranking the 24 problem sets according to the difference between the believable and unbelievable intuitive estimates. The first list then consisted of the believable, unbelievable, and neutral versions from ranked sets 1, 2, and 3, respectively. This cycle then continued through the remaining sets. The second list consisted of the unbelievable, neutral, and believable versions from sets 1, 2, and 3. The third list consisted of the neutral, believable, and unbelievable versions from sets 1, 2, and 3, respectively. Participants were assigned randomly to experimental lists. The 24 stimuli used in the main experiment are provided online (<http://blogs.umass.edu/rdcl/resources/>).

Procedure

The main experiment was presented on Ibx Farm through Mechanical Turk. On each trial, a Bayesian reasoning problem was shown as described above. The participant responded with an estimate of the posterior by typing a percent from 0 % to 100 %.

Results

All analyses were performed in R (R Core Team, 2016). A visual summary of the data is provided in Fig. 1. The left, middle, and right panels are data from the believable, neutral, and unbelievable conditions, respectively. The x -axis is the actual Bayesian posterior, and the y -axis is the mean of participants' estimated posterior. Each point represents 1 of the 24 problems in the relevant condition. The solid black line is the

regression line for these estimated posteriors. For reference, the dashed black line is the regression line for the mean intuitive estimates in each condition (points not shown).

Correct Bayesian inference would put all points on the gray diagonal. There are clear deviations from Bayesian inference in all conditions, but performance also differs across conditions. The proportion of variance in the mean estimated posterior for each problem that is accounted for by the actual posterior is 0.66 ($r = 0.81$) in the believable condition, 0.49 ($r = 0.70$) in the neutral condition, and 0.01 ($r = -0.11$) in the unbelievable condition. The relationship between the estimated and actual posteriors was significantly stronger in the believable than in the unbelievable condition ($z = 4.55$, $p < .001$) and was stronger in the neutral than the unbelievable condition ($z = 3.91$, $p < .001$; Diedenhofen & Musch, 2015, using z from Pearson & Filon, 1898).

To quantify this difference in another way, we calculated the mean absolute distance from the estimated posterior to the correct posterior for each subject, in each condition. These data are provided in Table 1 (i.e., $|p_{\text{est}}(H|D) - p(H|D)|$). There was an overall difference across conditions, $F(2, 106) = 16.29$, $p < .001$, $\eta_p^2 = 0.24$. Estimated posteriors for the believable stimuli were closer to the actual posterior than for the unbelievable stimuli, $t(53) = 4.92$, $p < .001$, $d = 0.61$, but were not significantly closer than for the neutral stimuli, $t(53) = 0.84$, $p = .41$, $d = 0.11$. The neutral and unbelievable stimuli also differed, $t(53) = 4.47$, $p < .001$, $d = 0.67$.

We next investigate how intuitive estimates factor into the judgment process. The simplest possibility is that participants ignored the rates and simply responded based on their intuitive estimates. The scatterplots in the top row of Fig. 2 show the relationship between the intuitive estimate (x -axis) for each problem and the mean of the estimated posterior judgments (y -axis). The believable, neutral, and unbelievable conditions are in the left, middle, and right, columns, respectively. The proportion of variance in the estimated posteriors accounted for by the intuitive estimates (R^2) is also provided in each panel. There is a high correlation between the intuitive estimate and the estimated posterior in the believable condition, presumably because these stimuli were selected intentionally so that the intuitive estimate was close to the actual posterior. However, there is little relationship

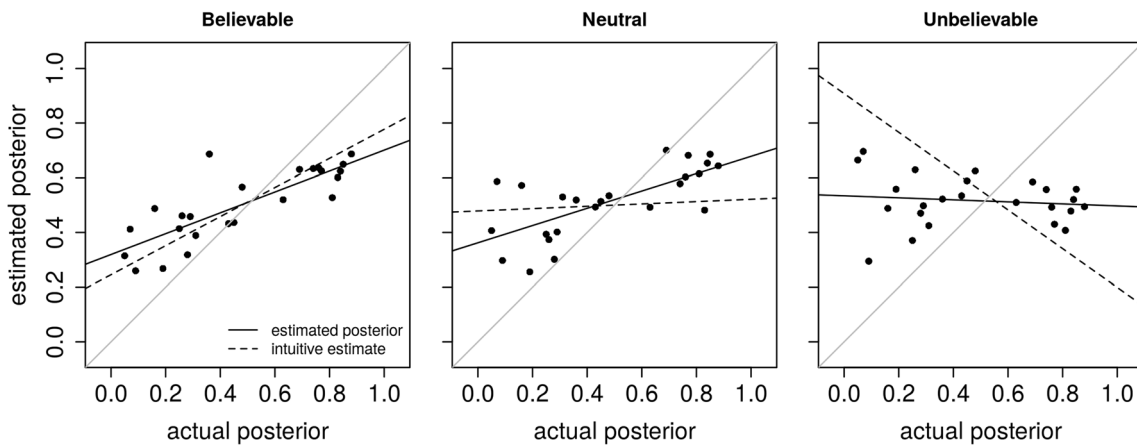


Fig. 1 Mean estimated posteriors plotted against actual posteriors in each believability condition. *Solid black line* Regression line for these points, *dashed black line* regression line for the intuitive estimates (points not shown)

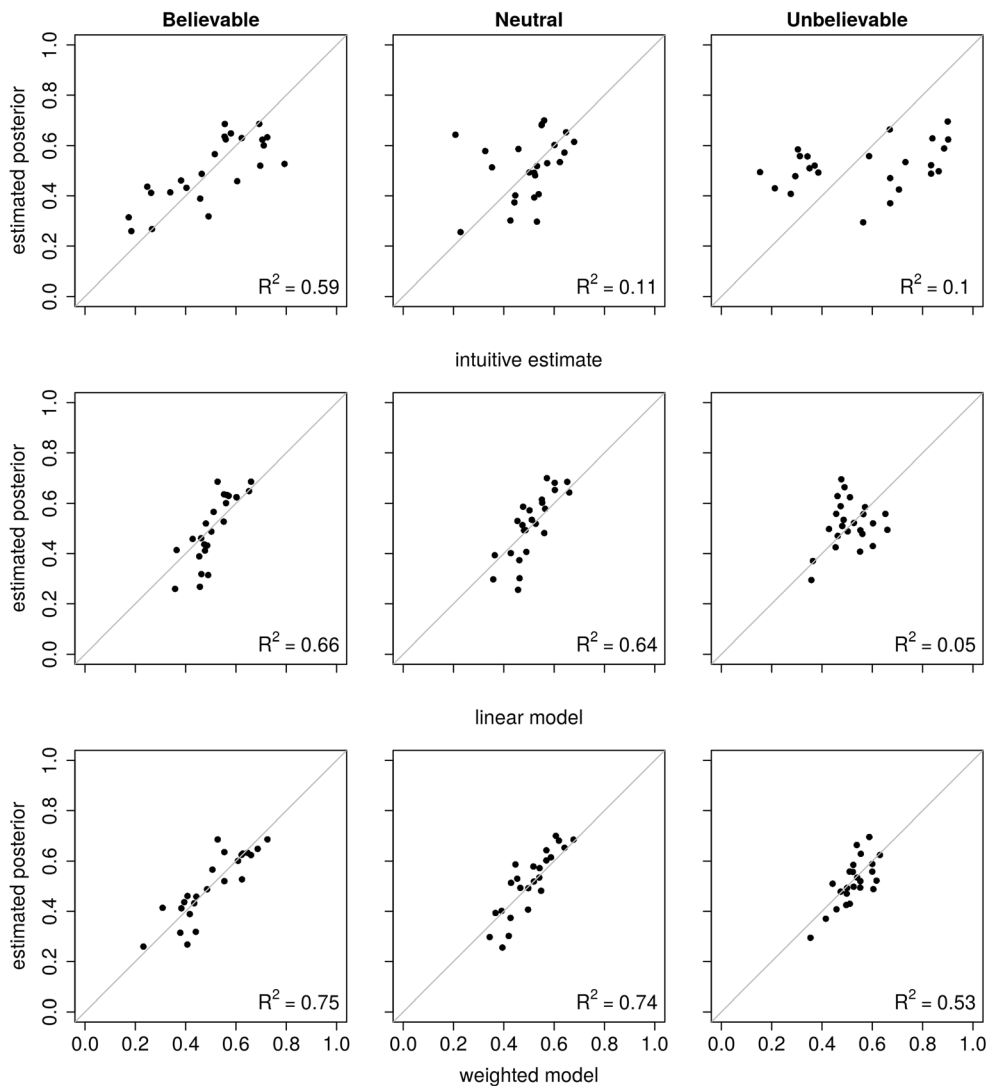


Fig. 2 Mean estimated posteriors plotted against intuitive estimates and predictions of the linear and weighted models in each believability condition. R^2 is provided separately in each panel, but the same parameters were used for all panels in a row

in the neutral and unbelievable conditions. Aggregating the three conditions, the data are not well accounted for by the intuitive estimates alone (overall $R^2 = 0.21$).

Clearly, the rates provided in each problem are likely to play some role in the reasoning process. In a similar task, Cohen and Staub (2015) found that, for many participants, the estimated posterior can be modeled as an additive function of these rates. First, we applied a simplified model that ignored individual variability, regressing the mean estimated posterior in each of the 72 problems on the three rates. Below we explore individual variability. The same four parameters (three rate coefficients and an intercept) were used for all three conditions. The predictions of this linear model are provided in the second row of Fig. 2. This model does a good job of accounting for the believable and neutral conditions, but not the unbelievable condition. Aggregating the three conditions, the overall $R^2 = 0.42$.

The final model combines the insights of the previous two models. The estimated posterior is modeled as a weighted combination of the linear model and intuitive estimates. There are five parameters (three rate coefficients, an intercept, and the weight) for the 72 data points. The results are provided in the third row of Fig. 2. This weighted model does an even better job of predicting both the believable and neutral conditions. Importantly, it can also now account for a large proportion of the variance in the unbelievable condition, and aggregating the three conditions, the overall $R^2 = 0.69$. This result is particularly impressive given the restricted range of the estimated posteriors in the unbelievable condition. The weights for the linear model and intuitive estimate were 0.68 and 0.32, respectively. Interestingly, including the *actual* posterior in this weighted model in place of the intuitive estimate provides no advantage over the linear model alone ($R^2 = 0.42$), and, in this model, the weight on the actual posterior was very close to 0. Thus, the influence of the three rates in the problem appears to be entirely captured by the linear model. Details of the models are provided online (<http://blogs.umass.edu/rdcl/resources/>).

Given prior evidence of individual differences in use of the rates (e.g., Cohen & Staub, 2015), these models were also fit separately to each individual's 24 observations, allowing all parameters to vary between individuals. The mean R^2 was 0.36 for the linear model and 0.45 for the weighted model that includes a term for the intuitive estimate. The weighted model better accounts for the 72 data points in the aggregated data than it does, on average, for the 24 data points generated by an individual participant. This result suggests that there is substantial noise in individual subject data, with the influence of the intuitive estimate emerging most strongly when participant estimates for each problem are averaged.

Further inspection of individual participant fits is revealing. Figure 3 is a histogram of the weights that the weighted model places on the linear combination of rates, for the 54

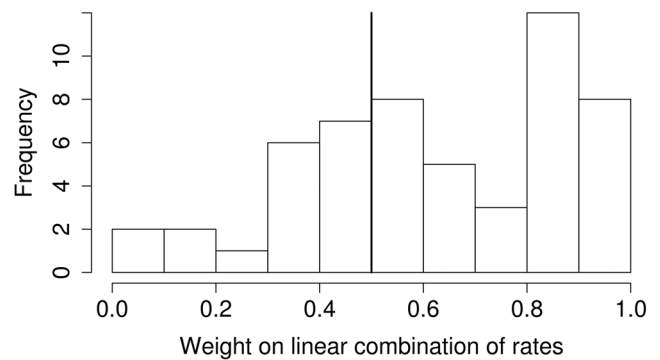


Fig. 3 Histogram of the weights on the linear combination of rates across individual participants

participants.² A value of 0 indicates that all weight was placed on the intuitive estimate, and a value of 1 indicates that all weight was placed on the linear combination of rates. A few clear patterns emerge. First, the mean is 0.68, which is identical to the weighting on the linear combination of rates in the averaged data. Second, a majority of participants (69 %) place more weight on the rates than the intuitive estimate. Third, the data are clearly bimodal: one group of participants places almost no weight on the intuitive estimate, while a second group shows a more even weighting of the rates and the intuitive estimate.

Figure 4 further illustrates individual variability in sensitivity to the intuitive estimate. On the x -axis is the fit of the linear model without the intuitive estimate, in R^2 , and on the y -axis is the fit of the weighted model that includes the intuitive estimate. When a participant's estimated posteriors are well modeled as a linear combination of rates, with $R^2 > .6$, there is no benefit at all from inclusion of the intuitive estimate term. Though some of this phenomenon is a simple ceiling effect (i.e., little improvement is possible if R^2 is near 1), it also suggests that those participants who do combine the rates in a consistent manner tend to ignore the intuitive estimate. On the other hand, many subjects who are poorly fit by the linear model show substantial improvement in fit when the intuitive estimate is included.

Discussion

This experiment provides evidence that real-world beliefs influence judgments in Bayesian reasoning problems. Although there is individual variability, when the normatively correct posterior is relatively close to the mean intuitive estimate, responses tend to be close to the correct answer; when the normatively correct posterior is far from this mean intuitive estimate, responses tend to be further from the correct answer.

² Due to local minima and the addition of jitter, some weights were slightly less than 0 or greater than 1. Weights have been capped at 0 and 1 in the Fig. 3.

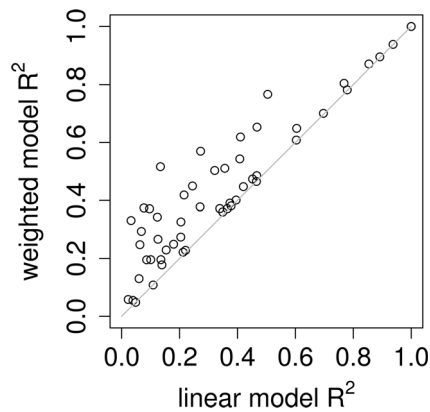


Fig. 4 Fit in R^2 of the full weighted model plotted against fit in R^2 of the model that includes only a linear combination of rates, for the 54 participants

Indeed, in the latter case, there was no relationship whatsoever between participants' estimated posterior and the correct posterior across a range of 24 problems (right panel of Fig. 1). A statistical model in which the mean estimated posterior in a given problem is a weighted sum of the mean intuitive estimate and an additive combination of the probabilities in the problem provides a much better fit to the data ($R^2 = .69$) than does a model based on either the intuitive estimate alone ($R^2 = .21$) or the linear combination alone ($R^2 = .42$). In particular, only the former model can account for the data when the actual posterior is unbelievable.

These results have implications for the interpretation of previous research using real-world scenarios as stimuli. The believability of the normatively correct posterior is likely to be a source of variability in how well subjects do in these problems, and therefore in how researchers evaluate reasoners' abilities. If an experiment were to make use only of problems like our unbelievable problems, the conclusion would likely be that people cannot do these problems at all, as in this case there is literally no relationship between the estimated posterior and the correct answer. On the other hand, if an experiment were to make use only of problems like our 'believable' problems, the conclusion would likely be that people are quite good at Bayesian reasoning, as the correlation between the correct answer and the estimated posterior is about 0.81. Neither of these conclusions would be warranted. Relatedly, experiments that contain a mix of believable and unbelievable problems may involve unanticipated item effects. Looking forward, the present results lead to the recommendation that researchers who are investigating Bayesian reasoning per se either take beliefs about real-world posteriors into account, or design problems with no real-world content (as in Cohen & Staub, 2015).

Turning to more theoretical implications, these results demonstrate that real-world beliefs contaminate abstract reasoning in a Bayesian context. The idea that two kinds of information are combined in the reasoning process has been previously

suggested for both logical (e.g., Beller & Spada, 2003; Evans, 2007; Evans & Stanovich, 2013) and Bayesian (e.g., Evans et al., 2002) reasoning. For example, a model that combines the output of decontextualized logical reasoning and prior knowledge provides a good account of a set of conditional reasoning results (Klauer, Beller, and Hutter, 2010). The study most similar to the present one is that by Evans et al. (2002), who explored the role of multiple sources of information in Bayesian reasoning. They found that personal beliefs can be used by participants when they are not supplied by the experimenter, and that these beliefs are weighted more heavily than provided information. Thus, both the Evans et al. (2002) results and the current results suggest that reasoners may discount information provided in Bayesian reasoning problems and instead rely on real-world beliefs.

It is important to note, however, that we diverge from Evans et al. (2002) in our underlying model of how reasoners perform these problems. Evans et al. (2002) assume that reasoners do perform essentially Bayesian reasoning, but with potentially non-normative weights on the rates. We have argued (Cohen & Staub, 2015) that reasoners combine the provided rates additively, and that most subjects actually take into account only one or two of the rates. In the present study we find yet another departure from normatively correct Bayesian reasoning, in which the posterior that is arrived at based on an additive combination of rates may be adjusted based on an intuitive estimate of that posterior. Normatively correct Bayesian reasoning plays no role in our model.

There are connections between the current emphasis on posterior intuitive estimate and work showing that knowledge of causal relations can influence probabilistic reasoning (e.g., Ajzen, 1977; Hayes, Hawkins, Newell, Pasqualino, & Rehder, 2014; Krynski & Tenenbaum, 2007; McNair & Feeney, 2015; Tversky & Kahneman, 1980). For example, telling participants that the presence of a cyst can cause a positive mammogram can improve performance on the standard medical diagnosis problem described above. One way to interpret the "unbelievability" of a posterior, then, is that a causal model may be lacking. For example, participants may not believe that a positive test is only mild evidence for breast cancer because they are not aware of the other factors that can cause a positive result. If unbelievability is understood in these terms, it is possible that providing a causal model results in an improvement in reasoning because it makes the correct posterior believable. That is, the believability of a posterior may be viewed as a matter of having an appropriate causal model, and the improvement in Bayesian reasoning performance that comes about when a causal model is provided may be viewed, in part, as an effect on the believability of the posterior.

Finally, we note that we have not addressed critical questions about exactly how real world beliefs about the posterior influence the reasoning process. If it is interpreted as a process

model, the weighted model assumes, in effect, that intuitive estimates do not influence the reasoning process per se, but rather bias the response. One component of the weighted model additively combines the three rates; the output of this component is weighted, and is combined with the intuitive estimate. Dube, Rotello, and Heit (2010) have made an analogous claim regarding the belief bias in syllogistic reasoning. Dube et al. (2010) argued that the tendency to judge a syllogism with a believable conclusion as valid is due to an effect of believability on the response criterion, rather than an effect on the perception of the argument's strength (cf., Trippas, Handley, & Verde, 2013). We acknowledge that further research is clearly needed to explore this issue in the context of Bayesian reasoning. It may turn out that, while the statistical model we have employed does a reasonably good job in explaining the variance in the data, it is inadequate as a process model of how intuitive estimates are actually used. Indeed, the fact that this statistical model does a better job at the aggregate level than when fitting individual participant data may be seen as circumstantial evidence in favor of this position. Regardless, the present research has demonstrated a tendency for reasoners to take into account beliefs about real-world conditional probabilities in Bayesian reasoning situations. This result adds yet another example to the catalogue of all-too-human departures from normative principles of reasoning.

References

- Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology*, 35, 303–314.
- Anderson, R., Leventhal, L., Fasko, D., Basehore, Z., Zhang, D., Billman, A., Gamsby, C., Branch, J., & Patrick, T. (2015). A form of belief bias in judgments of Bayesian rationality. Paper presented at the 56th annual meeting of the Psychonomic Society, Chicago, Illinois, USA.
- BarHillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44, 211–233.
- Beller, S., & Spada, H. (2003). The logic of content effects in propositional reasoning: the case of conditional reasoning with a point of view. *Thinking & Reasoning*, 9, 335–379.
- Cohen, A., & Staub, A. (2015). Within-subject consistency and between-subject variability in Bayesian reasoning strategies. *Cognitive Psychology*, 81, 26–47.
- Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, 10(4), e0121945. doi:10.1371/journal.pone.0121945
- Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, 117, 831–863.
- Eddy, D. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, England: Cambridge University Press.
- Evans, J. S. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove, UK: Psychology Press.
- Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11(3), 295–306.
- Evans, J. S. B. T., Handley, S. J., Over, D. E., & Perham, N. (2002). Background beliefs in Bayesian inference. *Memory & Cognition*, 30, 179–190.
- Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8, 223–241.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency Formats. *Psychological Review*, 102, 684–704.
- Hayes, B. K., Hawkins, G. E., Newell, B. R., Pasqualino, M., & Rehder, B. (2014). The role of causal models in multiple judgments under uncertainty. *Cognition*, 133(3), 611–620.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Klauer, K. C., Beller, S., & Hütter, M. (2010). Conditional reasoning in context: A dual-source model of probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(2), 298.
- Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, 136(3), 430.
- McNair, S., & Feeney, A. (2015). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychonomic Bulletin & Review*, 22(1), 258–264.
- Pearson, K., & Filon, L. N. G. (1898). Mathematical contributions to theory of evolution: IV. On the probable errors of frequency constants and on the influence of random selection and correlation. *Philosophical Transactions of the Royal Society of London, Series A*, 191, 229–311.
- R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL <https://www.R-project.org/>
- Trippas, D., Handley, S. J., & Verde, M. F. (2013). The SDT model of belief bias: Complexity, time, and cognitive ability mediate the effects of believability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(5), 1393.
- Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale, NJ: Erlbaum.