Decision

# A Bayesian Latent-Mixture Model Analysis Shows That Informative Samples Reduce Base-Rate Neglect

Guy E. Hawkins, Brett K. Hayes, Chris Donkin, Martina Pasqualino, and Ben R. Newell
Online First Publication, December 22, 2014. http://dx.doi.org/10.1037/dec0000024

Decision


2325-9965/14/$12.00 http://dx.doi.org/10.1037/dec0000024

# A Bayesian Latent-Mixture Model Analysis Shows That Informative Samples Reduce Base-Rate Neglect

Guy E. Hawkins, Brett K. Hayes, Chris Donkin, Martina Pasqualino, and Ben R. Newell
University of New South Wales

We examined the conditions under which sampling information from different probability distributions reduces base-rate neglect in intuitive probability judgments. To assess the impact of our manipulations, we employed a novel Bayesian latent-mixture model analysis that allowed us to quantify evidence for base-rate neglect. Experience with samples from the *posterior* distribution in the form of sequential sampling and a descriptive summary tally both markedly reduced base-rate neglect relative to baseline, and the summary tally improved performance over sequential sampling. Experience with samples from the *prior* distribution reduced base-rate neglect when conveyed as a descriptive summary, but not when sequentially sampled over time. The results indicate that (a) a summary of sample information can be more beneficial to judgment performance than sequentially sampling the same information, and (b) the benefits of sampling experience are more likely to be realized when the contents of the sample are perceived as directly relevant to the judgment problem. These findings help to clarify when and how sampling experience facilitates intuitive probability judgment.

*Keywords:* probabilistic judgment, base-rate neglect, sequential sampling, Bayesian analysis, belief updating

*Supplemental materials:* http://dx.doi.org/10.1037/dec0000024.supp



Since Tversky and Kahneman's seminal studies (Tversky & Kahneman, 1974), the ability to intuitively reason with probabilities and statistics has been held in relatively poor esteem (e.g., Gigerenzer & Gaissmaier, 2011; Newell, 2013). In large part, this conclusion is the result of holding probabilistic judgments to the normative standard of Bayes's theorem. Given hypothesis $H_j$ from a space of $J$ discrete hypotheses, Bayes's theorem gives the probability of hypothesis $H_i$ given data $D$ as

$$p(H_i \mid D) = \frac{p(D \mid H_i)p(H_i)}{\sum_j p(D \mid H_j)p(H_j)}$$

This formula contains three important components: the *prior probability*, $p(\mathrm{H})$, representing the degree of belief in a hypothesis before data have been observed; the *likelihood*, $p(D|\mathrm{H})$, representing the extent to which the data are consistent with the hypothesis; and the *posterior probability*, $p(\mathrm{H}|D)$, representing the degree of support the data afford to a particular hypothesis.

The Bayesian normative standard has traditionally been used to evaluate how people intuitively reason with probabilistic information. The prototypical paradigm presents participants with a cover story and a series of relevant statistics. One well-known example is the "mammogram problem" (cf. Eddy, 1982; Gigerenzer & Hoffrage, 1995; Krynski & Tenenbaum,

Guy E. Hawkins, Brett K. Hayes, Chris Donkin, Martina Pasqualino, and Ben R. Newell, School of Psychology, University of New South Wales.

This research was supported by the Australian Research Council Discovery Grant DP120100266 to the second and last authors. Some elements of the experimental data were presented at the 35th Annual Meeting of the Cognitive Science Society, 2013, Berlin, Germany, and were described in the conference proceedings. We thank Chris Moore, Ann Martin, and Kelly Jones for their assistance in data collection.

Correspondence concerning this article should be addressed to Guy E. Hawkins, School of Psychology, University of New South Wales, Sydney, NSW Australia, 2052. E-mail: guy.e.hawkins@gmail.com

2007), a version of which is shown in Figure 1. In this problem, three key statistics are presented: the prior probability, or base rate, of women in the population with breast cancer, $p(\mathrm{C}) = .01$, the likelihood or "hit rate" of the mammogram to detect breast cancer in women with cancer, $p(\mathrm{M}|\mathrm{C}) = .80$, and the "false-positive rate," $p(\mathrm{M}|\text{–C}) = .15$. These statistics allow calculation of the target quantity—the conditional probability of breast cancer given a positive mammogram—with Bayes's theorem,

$$p(C \mid M) = \frac{p(M \mid C)p(C)}{p(M \mid C)p(C) + p(M \mid -C)p(-C)} = \frac{.8 \times .01}{.8 \times .01 + .15 \times .99} \approx .051$$

Participants typically provide conditional probability estimates that are much higher than the normative solution, suggesting insufficient consideration of the low base rate. Performance on the mammogram problem is often cited as an instance of a more general bias of base-rate neglect in human judgment (cf. Eddy, 1982; Evans, Handley, Perham, Over, & Thompson, 2000; Gigerenzer & Hoffrage, 1995).

## Reducing Base-Rate Neglect Through Sampling

Many approaches have been proposed to shift intuitive judgments of probability toward more normative patterns of responding. For example, the natural frequency hypothesis suggests that presenting statistical information as frequencies (e.g., 8 out of 10 cases) rather than probabilities (e.g., .8 of cases) increases the rate of normative performance (e.g., Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). Alternative approaches suggest instructions clarifying set relations between the relevant samples (Barbey & Sloman, 2007; Evans et al., 2000) and provision of causal frameworks for relevant statistics (e.g., Krynski & Tenenbaum, 2007).

Another approach that has recently had considerable success in alleviating base-rate neglect and other judgment biases is to provide trial-by-trial sampling of the frequency of target events from relevant probability distributions (e.g., Fiedler, Brinkmann, Betsch, & Wild, 2000; Hogarth, Mukherjee, & Soyer, 2013; Hogarth & Soyer, 2011; Lejarraga, 2010; Sedlmeier, 1999). Hogarth and colleagues (Hogarth & Soyer, 2011; Hogarth et al., 2013), for example, suggested that people are less likely to neglect relevant statistics such as a low base rate following experience with the relevant sample. Previous research has largely focused on sampling from the most relevant conditional probability distribution: the posterior distribution – $p(\mathrm{C}|\mathrm{M})$. In the mammogram problem, sampling from the posterior distribution conserves the true rate of women with breast cancer from the subgroup of women who received positive mammograms. Posterior sampling improves rates of normative responding and recognition

**Mammogram problem**
Doctors often encourage women at age 50 to participate in a routine mammography screening for breast cancer.

From past statistics, the following is known:

1% of women had breast cancer at the time of the screening.
Of those with breast cancer, 80% received a positive result on the mammogram.
Of those without breast cancer, 15% received a positive result on the mammogram.
All others received a negative result.

Your task is to estimate the probability that a woman, who has received a positive result on the mammogram, has breast cancer.

*Suppose a woman gets a positive result during a routine mammogram screening. Without knowing any other symptoms, what are the chances she has breast cancer? ___%*

*Figure 1.* The mammogram problem used in the experiment.

of the normatively correct response (Fiedler et al., 2000; Hogarth & Soyer, 2011). For example, Hogarth and Soyer (2011) found that 17% of participants recognized the correct response in a multiple choice version of the mammogram problem, but following posterior sampling, this increased to 97%.

## "Kind" Experience

The beneficial effects of posterior sampling could arise from a range of mechanisms. Hogarth and Soyer (2011) conceptualize sample-based information existing on a continuum from *wicked* to *kind* experience. Wicked experience describes situations in which samples provide biased feedback, and kind experience describes tasks with clear structure and unbiased feedback. We argue that there are at least two components to kind experience that might facilitate performance: the presentation *format* of sample information and the *relevance* of the informational content of those samples.

**The format of sample-based information.** The presentation format of the samples refers to how participants obtain information about the distribution of positive and negative cases. Hogarth and Soyer (2011) stated, "across time, a person observes sequences of outcomes that can be used to infer the characteristics of the data generating process" (p. 435). Hogarth et al. (2013) also suggest that sequentially sampling from a relevant distribution may improve judgments because it builds on a well-established human capacity to encode sequentially presented frequency information (Zacks & Hasher, 2002). This implies that sequentially sampling outcomes from a probability distribution might provide unique benefits over other modes of obtaining sample-based information from the same distribution, such as a tally of sample outcomes.

It remains unclear, however, whether sampling experience per se is necessary to improve the accuracy of judgments under uncertainty. Previously observed improvements in judgment accuracy may have resulted from exposure to a representative distribution of positive and negative cases that accurately reflected the statistics given in the problem. If this is true then providing a *summary description* of a relevant sample distribution should produce an improvement in judgments under uncertainty similar to that found for sequential sampling.

For example, in the mammogram problem, those in the sampling condition may sample from the distribution of women with a positive mammogram and experience 19 cases without cancer and one with cancer. Those in the sample-summary condition would receive the same distributional information in tabular form without experiencing individual cases. If sequential sampling provides a unique benefit to understanding the sample outcomes of a data-generating process, as Hogarth and Soyer (2011) imply, then only those given this experience should show improved judgment accuracy relative to a baseline condition that only received a description of the problem (like Figure 1). If the statistical content of these samples is crucial, however, then both conditions should demonstrate improved judgment accuracy relative to a condition that just receives a description of the problem without sample-based information. To this end, in our experiment (described in detail below), we ensured that the sampling and sample-summary conditions received equivalent statistical information; summary tallies were yoked to the samples experienced by individuals in the trial-by-trial sampling condition (see Rakow, Demes, & Newell, 2008, for a related manipulation). Thus, any observed differences between the two conditions must be due to the presentation format.

**The relevance of sample-based information.** The relevance of a sample refers to the specificity of the information contained in the samples to the judgment task. Sampling from the posterior distribution in previous research is arguably the most relevant experience that one could receive, as those samples approximate the true (normative) answer with increasing sample size. In fact, posterior sampling altogether removes the need to integrate the various components of Bayes's theorem, allowing the participant to obtain the normative answer from the sample information alone. A less relevant form of experience might allow participants to sample from the prior distribution (e.g., the population of women). Such sampling highlights the low frequency of the base-rate event (e.g., women with breast cancer), and is therefore more helpful than receiving no sample-based information, but will not approximate the normative answer with increasing sample size. In-

deed, if the phenomenon referred to as base-rate neglect actually reflects a neglect of the base-rate statistic, then we should observe improved performance when we increase the salience of the base-rate statistic with sampling experience.

If sampling experience highlights previously neglected outcomes in the sample space then the sufficient condition for improvements to judgment accuracy is a sampling distribution in which the base-rate event is rare. Accordingly, experience with the prior distribution (e.g., the population of women with a 1% base rate of cancer) should lead to similar facilitation in probability estimates to sampling from the posterior distribution, in which both conditions should provide estimates closer to the normative solution than a baseline description condition. However, if participants are sensitive to the specific content of the observed samples and their relevance to the judgment task, rather than the relative frequency of the target event (cf. Fiedler, 2008), then estimates should shift toward more normative responses when sampling from the posterior distribution, but not necessarily when sampling from the prior distribution. This occurs because samples from the prior distribution do not provide the "answer" to the target problem in the same way as samples from the posterior distribution. We tested the impact of format and relevance of sample information on base-rate neglect in our experiment.

## Issues in the Analysis of Base-Rate Neglect Problems

In many previous studies of base-rate neglect, the distribution of probability estimates has suggested that there are at least two discrete types of respondents: those who neglect the base rate and those who do not (cf. Bar-Hillel, 1980; Cosmides & Tooby, 1996; Evans et al., 2000; Gigerenzer & Hoffrage, 1995; Krynski & Tenenbaum, 2007; see Figure 2 below). In the mammogram problem, this typically manifests as those participants who neglect or incorrectly use the base rate giving high estimates, and those who do not neglect the base rate (whether they perform normative calculations or not) providing low estimates, with few intermediate estimates. In problems like the one shown in Figure 1, low estimates are generally closer to the normative solution, so the researcher aims to determine whether a particular experimental manipulation increases the number of "low-estimate respondents." This is a difficult statistical problem using conventional approaches that rely on location measures like means or medians (such as analysis of variance), as the presence of more than one latent respondent class can lead to bimodal response distributions (Lantz, 2013).

A survey of the literature reveals a number of approaches that have been developed to deal with data from multiple classes of responders. One common method is to attempt an ad hoc classification of participants into a variety of response categories (e.g., correct Bayesian, base-rate neglect, likelihood neglect) based on their subjective probability estimates. The counts of participants in these categories are then analyzed with nonparametric tests, such as a chi-square test, to assess the impact of some experimental manipulation (e.g., Evans et al., 2000; Krynski & Tennenbaum, 2007; McNair & Feeney, 2014a, 2014b). Such nonparametric tests are riddled with statistical issues, such as low power and their own set of assumptions, which are often violated in experiments of this type (Milligan, 1980).

A potentially more troublesome aspect of current approaches is that they require two steps. First, researchers observe their data and then subjectively decide on an appropriate cutoff point. Then, inference is performed on differences in the proportion of participants in each group across experimental conditions. Two serious issues arise here. First, the inference carried out in the second step assumes that the cutoff point was determined a priori (i.e., group classification is treated as an independent variable). Second, it was assumed that the chosen cutoff point was the only possible way in which the data could be divided. In what follows, we proposed a method that deals with all of the above issues. In particular, our approach used the observed data to simultaneously infer both the appropriate cutoff point and any influence of experimental manipulations on the proportion of participants in each class of respondents.

It is worth noting that the problem of analyzing discrete populations of respondents is not isolated to the study of base-rate neglect. Any experiment in which participants use different

strategies to solve the problem that the researcher imposes on them can yield multiple latent classes of respondents.

### Analyzing base-rate neglect with Bayesian mixture models

We identified discrete, latent classes of responders using a Bayesian mixture model (for introduction see Bartlema, Lee, Wetzels, & Vanpaemel, 2014). In our approach, we simply assumed that there were two general ways in which participants could answer the mammogram problem. The data were then used to infer the properties of the two groups, such as the average probability estimate in each group. As such, we predicted that the low- and high-estimator types would emerge from the model, because they were present in the data. At the same time, it is important to note, the model would also estimate the proportion of participants who fell into the category of low and high responders.

As in most studies, our aim was to determine whether an experimental manipulation would influence the number of people who gave low estimates (i.e., closer to a normative response in the mammogram problem). Hence, the proportion of low-estimate responders in a given experimental condition was our primary outcome measure. Our approach permitted us to compare the value of this mixture proportion across conditions with Bayesian hypothesis testing, which provides many benefits (e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). In particular, our approach allowed a principled method to directly compute the relative evidence for the null and alternative hypotheses.

The most important property of the Bayesian mixture model is that it is a one-step procedure. That is, we estimated the proportion of low-estimate participants across all experimental conditions simultaneously, at the same time that the properties of the low- and high-estimator groups were inferred. All inferences we made about differences across experimental conditions took into account the uncertainty surrounding classification of individuals into two groups of responders. As a result, our approach did not suffer from the issues associated with the standard two-step procedure. We did not treat class membership like an independent variable, and our inference did not assume a single possible classification.

## Method

### Participants

Undergraduate psychology students ($N$ = 175, $M_{age}$ = 20.1 years, $SD_{age}$ = 5.2) from the University of New South Wales participated for course credit. All were tested individually.

### Design

The experiment had five conditions manipulated between subjects: description-only, posterior sampling, posterior-sample summary, prior sampling, and prior-sample summary. The experiment was necessarily conducted in two phases, and from the outset we planned to recruit 25 participants per cell. Fifty participants were first randomly assigned to the description or posterior-sampling conditions ($n$ = 25 per cell). Because the sample summaries were yoked to the sample outcomes observed in the sampling condition, the posterior sample-summary condition ($n$ = 25) was conducted after the posterior-sampling condition. We collected data in the prior sampling and prior sample-summary conditions at a later date ($n$ = 25 per cell), again in a two-phase design to yoke the prior sample summaries to the observed sample outcomes in the prior-sampling condition. We later added another 25 participants to the prior-sampling and prior sample-summary conditions to ensure that we had enough precision in our measurement for a total of 50 participants per cell in these two conditions. All participants were recruited from the same undergraduate population in the same way and we observed no demographic differences between samples (e.g., age, sex, native language, numerical ability). In what follows, we analyzed all conditions together. We did not collect data from any additional experimental conditions or manipulations that are not reported in this study.

### Procedure

All participants were presented with the mammogram problem shown in Figure 1 (cf. Krynski & Tenenbaum, 2007, Experiment 2). In all conditions, the problem description (the no-

nitalicized text in Figure 1) was first presented on a computer screen.

In the description condition, an open-ended question requesting an estimate of the probability of cancer in a woman with a positive mammogram appeared after 15 s. The format of this estimate was a percentage chance of cancer between 0 and 100. Participants were invited to use an on-screen calculator to assist in solving the problem.[1]

Participants in the sampling conditions received an additional sampling phase between the problem description and the request for a probability estimate. In the posterior-sampling condition, participants were told that they could observe samples of women who had received a positive mammogram (i.e., "In order to assist you in this task, you will now be able to use a simulator to 'meet' a series of women, all of whom have received a positive mammogram;" for details of a similar procedure, see Hogarth & Soyer, 2011). Formally, this involved sampling from the posterior distribution, $p(C|M)$, such that each time the participant clicked a "simulate" button, with probability .051 (i.e., the approximate posterior solution) the participant was told "this woman has breast cancer," otherwise they were informed that "this woman does not have breast cancer." In the prior-sampling condition, participants were told they could observe samples of women who were about to undergo mammography screening (i.e., "In order to assist you in this task, you will now be able to use a simulator to 'meet' a series of women, all of whom are at age 50 and about to undergo routine mammography screening"). Each time the participant clicked a "simulate" button, with probability .01 (i.e., the cancer base rate) the participant was told that "this woman has cancer," otherwise they were informed that "this woman does not have cancer."

In both sampling conditions there was no limit on the number of samples that could be drawn. At any time during sampling, participants could also click an on-screen button to view a running tally of (a) samples with cancer, (b) samples without cancer, and (c) total samples viewed. The summary tally equated memory load between the sampling and sample-summary conditions (described below).[2] To familiarize participants with the sampling tool, prior to commencing the main experiment, they were shown the outcomes of 15 samples of tossing an unbiased coin. Following the sampling phase, participants were presented with the same open-ended question as the description condition.

The procedure for the sample-summary conditions was similar to the description condition, with the exception that participants were provided with an on-screen tally of sample outcomes from the relevant distribution. In the posterior sample-summary condition, participants were shown positive and negative cases of cancer from samples of women who had received a positive mammogram. Twenty-five tallies were generated based on the observed sample outcomes from the posterior-sampling condition. Participants in the prior sample-summary condition were presented with a summary tally of positive and negative cases of cancer from samples of women who were 50 years of age and about to undergo routine mammography screening, yoked to the prior sampling information. Fifty such individual tallies were constructed based on the samples drawn by participants in the prior-sampling condition.

Finally, participants completed a computer-based version of the four-item Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal

[1] An on screen calculator was provided in the description and sample-summary conditions. In those cells, probability estimates did not differ between those who used the calculator (slightly more than half of the participants) and those who did not. An aggregate Bayes factor provided evidence for the null hypothesis of no difference between these groups ($BF_{01} = 6.89$).

[2] Participants in the posterior sampling and prior-sampling conditions could access a summary tally throughout the sampling phase. There was a slight trend for participants in the posterior-sampling condition that utilized the summary (slightly more than two thirds of participants) to provide lower probability estimates, but this trend was reversed for participants in the prior-sampling condition. An aggregate Bayes factor provided indeterminate evidence for a difference between tally users and nonusers ($BF_{01} = .41$). To confirm that access to a summary tally did not influence the pattern of results reported in the main text, we reran the posterior-sampling condition with identical methods as reported in the main text but without access to a summary tally ($n = 25$). There was some evidence for the null hypothesis that the presence (or absence) of a summary tally did not change the proportion of low estimators when sampling from the posterior distribution of the judgment problem ($BF_{01} = 3.11$).

& Garcia-Retamero, 2012). After providing probability judgments and before the numeracy test, the cancer estimation question was repeated together with four alternative "answers that people commonly give to this question" (1%, 5%, 65%, 80%). Participants responded with a mouse click to the option they thought was "closest to the correct answer." Rates of choice of the approximately correct answer (5%) showed similar patterns of differences between groups as the probability estimates, so we do not report them further. No additional variables were recorded and there was no time limit on any part of the procedure.

## Results

### Preliminary Analyses

Numerical ability as measured by the Berlin Numeracy Test (Cokely et al., 2012) was similar across the five groups: description ($M = 2.32$ out of a possible 4, $SE = .24$), posterior-sampling ($M = 2.40$, $SE = .23$), posterior sample-summary ($M = 2.48$, $SE = .25$), prior-sampling ($M = 2.52$, $SE = .16$) and prior sample-summary conditions ($M = 2.46$, $SE = .15$). These data were analyzed using the approach developed by Morey and Rouder (2013), who implemented common significance tests such as ANOVA within a Bayesian framework. This approach produced Bayes factors for linear model effects that indicated the weight of evidence for or against the null hypothesis, directly interpretable to how many times more likely one hypothesis was to have generated the observed data over another. We use the notation $BF_{01}$ to refer to Bayes factors, where $BF_{01} > 1$ indicates support for the null hypothesis and $BF_{01} < 1$ support for the alternative hypothesis. For example, $BF_{01} = 10$ indicates that the data are 10 times more likely to have come from the null hypothesis than the alternative hypothesis, and $BF_{01} = .1$ indicates the opposite conclusion. This analysis revealed no evidence of group differences in numeracy, $BF_{01} = 41.18$.

We also examined behavior in the sampling conditions. In the prior-sampling condition, participants sampled an average of 21.5 cases ($SE = 2.7$, range = 2–84) from the prior distribution (i.e., "met" women who were about to undergo mammography screening), compared with 16.1 cases ($SE = 2.5$, range = 3–50) in the posterior-sampling condition, $BF_{01} = 1.95$. The mean proportion of observed cancer samples did not reliably differ from the expected proportion of cancer samples as defined by the prior probability, $p(C) = .01$, or the posterior solution, $p(C|M) \approx .051$, $BF_{01} = 4.74$ and $BF_{01} = 3.48$, respectively. This meant that the mean proportion of observed positive cancer cases was larger in the posterior- than the prior-sampling condition ($M = .054$, $SE = .015$ vs. $M = .01$, $SE = .003$, respectively), $BF_{01} = .01$. As expected given the low prior probability, the majority of participants in the prior-sampling condition (78%) never observed a positive cancer case, and the remainder observed a single positive cancer case, whereas 56% of participants in the posterior-sampling condition observed at least one positive cancer case. The ratio of participants that observed at least one positive cancer case was lower in the prior-sampling condition (yes = 11, no = 39) compared with the posterior-sampling condition (yes = 14, no = 11), $BF_{01} = .06$. The sample characteristics of the posterior and prior-sampling conditions suggest that the manipulations successfully demonstrated differential rates of positive cancer cases in the observed samples.

### Probability Judgments

Probability estimates are shown in the upper row of Figure 2 as violin plots, a combination of a box plot and a kernel-density estimate to convey information about the distribution of the data. The violin plots confirm that there was considerable bimodality in probability estimates. Regardless of condition, participants generally provided high estimates (e.g., 60–90%) or low estimates (below 20%) with few intermediate values. This is clearly illustrated when estimates are collapsed across conditions, shown in the vertical histogram at the upper right of Figure 2. The most striking features of the violin plot are that (a) the prior conditions did not experience the same degree of benefit of sampling information as did the posterior conditions, and (b) the sample-summary conditions had

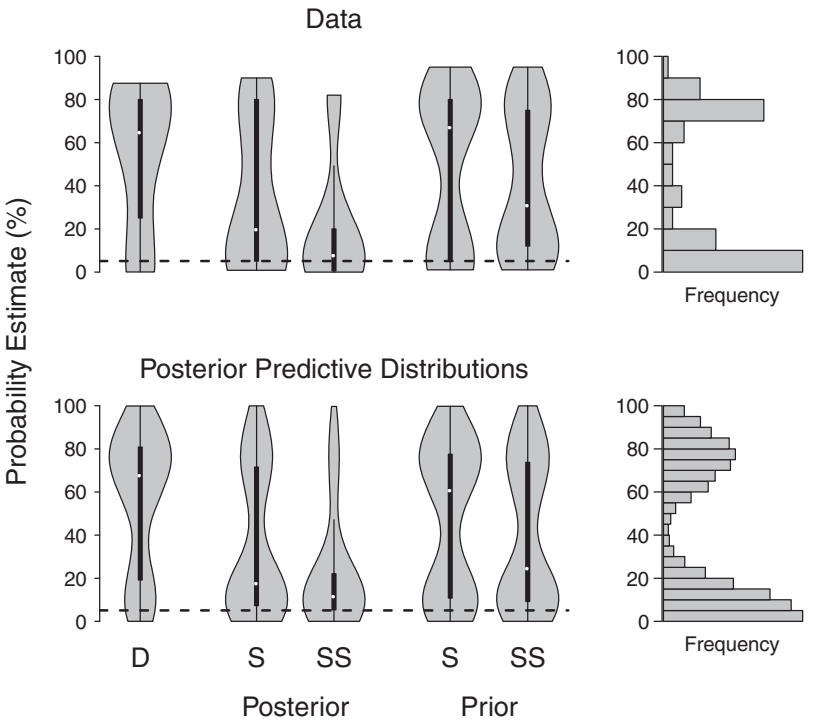


*Figure 2.* Violin plots of probability estimates in data (upper row) and posterior predictive distributions (lower row) as a function of the sampling distribution in the judgment problem (posterior, prior) and information format (D = description, S = sampling, SS = sample summary). The dashed horizontal line represents the normative solution. Histograms on the right illustrate the distribution of estimates collapsed across the five conditions, separately for data and posterior predictive distributions. Each 'violin' combines a boxplot and a kernel density estimate. The boxplot component is indicated with the white circular symbol (median), the interquartile range (heavy vertical line), and 1.5 × interquartile range (thin vertical line) as an indicator of the range of scores. The 'violin'-like shape of each distribution is obtained through a smoothed density estimate of the data, rotated vertically, and plotted on both sides of the box plot to create a symmetric figure. The width of the violin is proportional to the number of data points that fall in that part of the distribution.

improved performance (i.e., lower probability estimates) relative to the sequential sampling conditions.

## Bayesian Model-Based Analysis of Probability Estimates

Our latent-mixture model assumed that there were two populations of respondents in the mammogram problem—low and high estimators. Low estimators will generally be closer to the normative solution for this problem. Hence, the proportion of low-estimate responders in each experimental condition is the primary dependent measure, where a greater proportion of low estimators indexes better performance. Full details of the model are given in Section A of the online supplemental material.

The Bayesian model captured the qualitative and most quantitative aspects of participants' probability estimates, as shown in Figure 2. The close correspondence between model predictions and probability-estimate data indicated that we could safely interpret the model parameters.

**Proportion of low estimators.** Figure 3 shows posterior distributions for the proportion of low estimators in each condition. The proportion of low estimators increases as one moves from the description condition to the

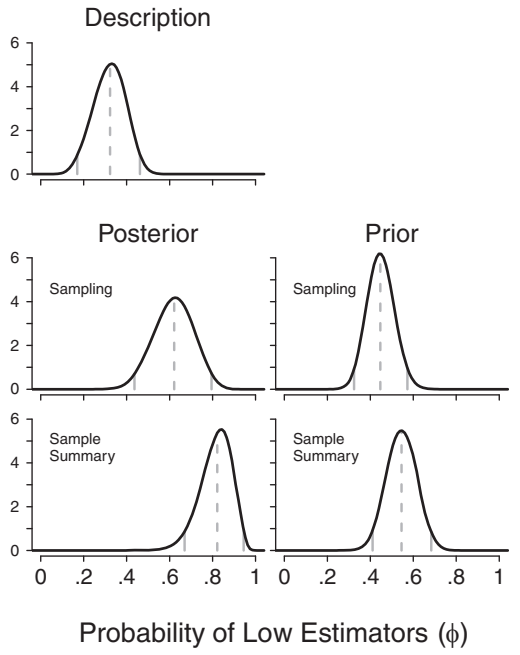


*Figure 3.* Posterior distributions for the probability of a response from the low-estimator distribution (ϕ) in the experiment. Rows indicate experimental conditions: baseline description (top), the participants who sequentially sampled outcomes from the posterior distribution in the judgment problem (middle, left), the prior distribution in the problem (middle, right), and participants who received a summary tally of sample outcomes from the posterior distribution in the problem (lower, left), or the prior distribution in the problem (lower, right). Dashed gray lines indicate the median of the marginal posterior distribution of parameters and the pair of solid gray lines in each panel represent the lower and upper bounds of the 95% highest density interval (Kruschke, 2011)—the smallest interval required to contain 95% of the posterior density for a parameter. The degree of overlap between any two marginal posterior distributions of parameters indicates the extent to which those two distributions contain the same estimate for the proportion of low estimators.

posterior-sampling and posterior sample-summary conditions. In contrast, the posterior distributions for the prior-sampling and prior sample-summary conditions are approximately equal, with some suggestion of a greater proportion of low estimators in the prior sample-summary condition.

**Bayesian hypothesis tests.** To test whether the sampling and sample-summary conditions performed better than the description condition, we conducted one-sided Bayesian hypothesis tests using the Savage–Dickey density ratio test (for a tutorial see Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). The Savage–Dickey density ratio gives a Bayes factor indicating the degree to which the data are likely to have come from one of two nested models. We refer the reader to Section B of the online supplemental material for details of the Savage–Dickey density ratio and our approach to hypothesis tests.

Experience with the posterior distribution via sampling experience *or* a sample summary reliably increased the proportion of low-estimate respondents relative to the baseline description condition, as shown in Figure 4 (i.e., $BF_{01}$ = .026 and $BF_{01}$ = .0001 for the sampling and sample-summary conditions, respectively). In contrast, sampling from the prior distribution did not reliably increase the proportion of low estimators relative to the description condition, $BF_{01}$ = .474. However, receiving a sample

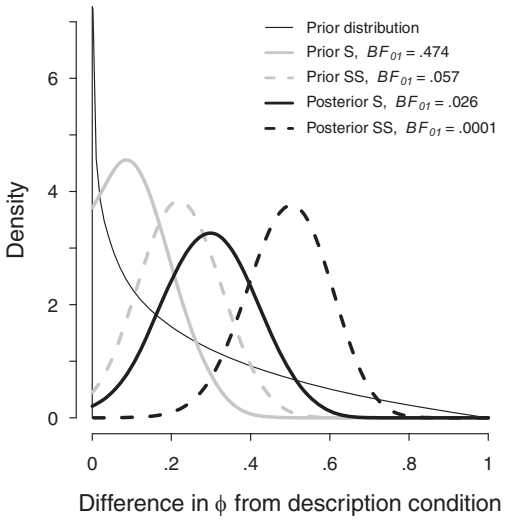


*Figure 4.* Posterior distributions of the one-sided test that the four experimental conditions had a greater proportion of responses from the low estimator distribution (ϕ) than the description condition. The thin solid black line shows the prior distribution on the one-sided hypothesis in the Bayesian hypothesis tests. Black and gray lines show the participants whose samples were drawn from the posterior distribution in the judgment problem, and those whose samples came from the prior distribution in the problem, respectively. Solid and dashed lines show the sampling (S) and sample-summary (SS) conditions, respectively. Bayes factors for the one-sided hypothesis tests are shown in the legend; values $< 1$ indicate support for the alternative hypothesis. See main text and Section B of the online supplemental material for full details.

summary from the prior distribution improved performance compared with the baseline description task, reflected as a greater proportion of low estimators, $BF_{01} = .057$.

We also tested whether the proportion of low-estimate responders differed between the sampling and sample-summary conditions. A two-sided test that allowed for the possibility of increased or decreased performance in the sampling compared with sample-summary conditions indicated that the posterior sample summary and posterior-sampling conditions were approximately 7 times more likely to have a different (rather than the same) proportion of low-estimate responders ($BF_{01} = .137$). The lower left panels of Figure 3 suggest that there was superior performance in the posterior sample-summary compared with the posterior-sampling condition. Although Figure 4 suggests that the proportion of low estimators was larger in the prior sample summary than in the prior-sampling condition, the evidence for this effect was not convincing, $BF_{01} = .423$.

It was clear that when presentation format was equated, the *relevance* of the distribution to the judgment task affected the proportion of low estimators. There was moderate support that the posterior-sampling condition had a greater proportion of low estimators than the prior-sampling condition, $BF_{01} = .197$ (solid black and gray lines in Figure 4). There was strong evidence that judgments were more accurate in the posterior-sample summary compared with the prior sample-summary condition, $BF_{01} = .025$ (dashed black and gray lines in Figure 4).

## Discussion

We examined the conditions under which sampling information leads to improvements in intuitive probability judgments. Consistent with previous research (e.g., Hogarth & Soyer, 2011), we found that sequential sampling from the posterior distribution reduces base-rate neglect compared with a baseline description condition. However, we have shown that this facilitation is not a function of sampling experience per se. Provision of the same information in a brief sample summary led to even better performance. Moreover, we found that the source of the sample is important. Sampling from the prior distribution of the judgment problem can be seen as less relevant to the task of estimating a conditional probability and was found to produce smaller facilitation than information from the posterior distribution of the judgment problem. A sample summary of the prior distribution did, however, reduce base-rate neglect relative to the description baseline.

All of our conclusions were drawn from a Bayesian model-based analysis and Bayesian hypothesis tests. Our latent-mixture model approach is a substantial improvement over previous approaches to examine discrete differences between individuals because it allows for uncertainty in the parameter values for the two classes of responders (the means of the low and high estimators) and the cutoff point between groups (the mixture probability, $\phi$). The analysis produced an outcome measure of direct interest to the research question: the proportion of participants that performed "well" in the task while simultaneously inferring what constituted a low or high response. The value of this mixture proportion can then be compared across experimental conditions with Bayesian hypothesis tests, which offer a principled method for assessing evidence for equivalence between groups of interest (i.e., for the null hypothesis), as well as for group differences.

### The Relevance, Content, and Format of Samples

Finding that the source of the sample is important supports a thesis based on highly relevant sample information: Provision of samples that highlight the rarity of the target outcome *and* are immediately relevant to the problem solution are more likely to reduce base-rate neglect. Samples from a posterior distribution meet this requirement. This contrasts with sample information that highlights the low frequency of a rare outcome, but is not drawn from the relevant, conditionalized distribution. Samples from the prior distribution fall in this latter category, and they only appear to affect base-rate neglect when presented in a simplified summary format. This result also suggests that our participants were not completely myopic with respect to the source and relevance of the sample, thus contrasting with some, perhaps overly pessimistic assessments of people's metacognitive awareness regarding reliance on samples (e.g., Fiedler, 2012).

Our results therefore suggest that the "kindness" of experience is not tied to a particular format of information presentation or acquisition. Rather, the key factor in improving intuitive probability judgments is the observation of a relevant and representative distribution of positive and negative cases (cf. Rakow et al., 2008): Facilitation might arise from helping participants represent the various statistics in usable ways for solving a problem.

There was generally a greater reduction in base-rate neglect when participants received a summary of sample outcomes rather than having experienced sequentially sampled outcomes, though this effect was stronger in the posterior conditions. This finding contrasts with Hogarth and Soyer's (2011) claim that sequential experience provides unique benefits for learning and representing statistical information. In subsequent work with judgment problems that did not include the mammogram task, Hogarth et al. (2013) found that judgment accuracy was better following a combination of sequential sampling and access to summary tallies than sequential sampling alone. We did not find a corresponding effect in our experiments[2]. This difference may be due to the nature of the summary tallies: Hogarth et al. (2013) provided participants with an always-visible summary tally that was updated following each sample, whereas our participants were required to click a button to view a summary tally. Hogarth et al. (2013) interpreted their result as evidence that summary tallies reduced the load on memory of aggregating the outcomes of sampling trials. Our findings go further by showing that summaries of representative samples without sampling experience can facilitate probability judgments.

It is possible that the approximately ordinal effect of information format that we observed, in which the sample-summary conditions outperformed the sequential sampling conditions, which in turn generally outperformed the description condition, was due to the relative salience of the available information (cf. Fantino & Navarro, 2012). Analogous to the description condition, in the absence of modifying (sample) information, participants make predictions that are consistent with their model of the environment that breast cancer leads to a positive mammogram (Kahneman & Tversky, 1972). This pattern of responding is only gradually eroded with sampling experience (e.g., Newell & Rakow, 2007), which suggests that the sequential sampling conditions responded with a combination of the described information and the (normatively irrelevant) sample-based information. In turn, a sample summary is more salient in the stimulus display than the described probabilities, resulting in greater attention toward the summary tally and response patterns consistent with the information they contain.

It is important to note that the facilitation brought about by sample-based information does not necessarily lead to the *correct* answer—even our low estimators tended toward overestimation—but having access to a relevant, representative sample appears to disabuse participants of the notion that a positive result on a mammogram test necessarily leads to a high probability of breast cancer. This conclusion emerged when the statistics and cover story were held constant across manipulations of the sampling distribution and information format. There are a number of reasons to believe that our results are likely to generalize to other statistics and cover stories. Base-rate neglect is observed under standard conditions (i.e., our description condition) in problems with various cover stories that use a range of low and high base-rate statistics (for a recent example, see McNair & Feeney, 2014a). Furthermore, Hogarth and Soyer (2011) found that sampling experience improved performance with the same low base rate as our problem (1%), but with different likelihood and false-positive statistics, which suggests some generalizability of the sampling-facilitation effect. Finally, in related paradigms from the judgment literature in which nonoptimal patterns of responding are common, such as probability matching and the sunk-cost effect, the addition of described information to experienced samples can lead to more optimal patterns of responding (e.g., Fantino & Navarro, 2012; Newell, Koehler, James, Rakow, & van Ravenzwaaij, 2013), as observed in our posterior sequential sampling condition.

With respect to information format, there are a number of reasons our findings cannot be explained away as simply another example of the "natural" advantage of frequency presentations over probabilistic formats (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). We found a facilitation effect even though we only provided frequency information about a

single distribution; the remaining statistical information (e.g., false-positive rates) in the problem was presented in probability format. Sample summaries did not correspond to large sample sizes—so judgments could have been influenced by the law of small numbers—or necessarily led to easily computed integer solutions to the conditional probability problem. Moreover, although those in the sampling conditions were exposed to frequency information, their responses were required in percentage format. According to the natural frequency view, improvements in judgment should only be found when both relevant statistics and the intuitive estimate are framed in frequency formats (cf. Evans et al., 2000).

## Conclusion

Our results show that the positive effects of sampling on probabilistic judgments are influenced by two factors. First, the samples must be perceived as relevant to the target judgment. Second, sampling experience per se seems less important than having a summary of representative samples. In this sense, the experience of "flipping a coin" can shift people toward a better understanding of the relative probability of alternative outcomes, but only when flipping the "right" coin (one that yields a representative sample from the target distribution). Performance can also be facilitated without seeing the sequence of coin flips, and under some circumstances, a detailed summary of sample outcomes can lead to more accurate estimates than sequential experiences. More generally, our results highlight when and where sampling experience might help judgment performance in base-rate neglect problems, and demonstrate that sampling experience per se is not a panacea. Moreover, this work illustrates a technique for analyzing subjective probability estimates that has many advantages over existing methods, and leads to more robust conclusions about the effects of experimental manipulations.

## References

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences, 30,* 241–254. http://dx.doi.org/10.1017/S0140525X07001653

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44,* 211–233. http://dx.doi.org/10.1016/0001-6918(80)90046-3

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology, 59,* 132–150. http://dx.doi.org/10.1016/j.jmp.2013.12.002

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making, 7,* 25–47.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58,* 1–73. http://dx.doi.org/10.1016/0010-0277(95)00664-8

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511809477.019

Evans, J. S. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition, 77,* 197–213. http://dx.doi.org/10.1016/S0010-0277(00)00098-6

Fantino, E., & Navarro, A. (2012). Description–experience gaps: Assessments in other choice paradigms. *Journal of Behavioral Decision Making, 25,* 303–314. http://dx.doi.org/10.1002/bdm.737

Fiedler, K. (2008). The ultimate sampling dilemma in experience-based decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 186–203. http://dx.doi.org/10.1037/0278-7393.34.1.186

Fiedler, K. (2012). Chapter one: Meta-cognitive myopia and the dilemmas of inductive-statistical inference. *Psychology of Learning and Motivation, 57,* 1–55. http://dx.doi.org/10.1016/B978-0-12-394293-7.00001-7

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base-rate neglect and statistical format. *Journal of Experimental Psychology: General, 129,* 399–418. http://dx.doi.org/10.1037/0096-3445.129.3.399

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62,* 451–482. http://dx.doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704. http://dx.doi.org/10.1037/0033-295X.102.4.684

Hogarth, R. M., Mukherjee, K., & Soyer, E. (2013). Assessing the chances of success: Naïve statistics versus kind experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 14–32. http://dx.doi.org/10.1037/a0028522

Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: Kind experience versus nontransparent description. *Journal of Experimental Psychology: General, 140,* 434–463. http://dx.doi.org/10.1037/a0023265

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454. http://dx.doi.org/10.1016/0010-0285(72)90016-3

Kruschke, J. K. (2011). *Doing Bayesian data analysis*. Waltham, MA: Academic Press.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136,* 430–450. http://dx.doi.org/10.1037/0096-3445.136.3.430

Lantz, B. (2013). The impact of sample nonnormality on ANOVA and alternative methods. *British Journal of Mathematical and Statistical Psychology, 66,* 224–244. http://dx.doi.org/10.1111/j.2044-8317.2012.02047.x

Lejarraga, T. (2010). When experience is better than description: Time delays and complexity. *Journal of Behavioral Decision Making, 23,* 100–116. http://dx.doi.org/10.1002/bdm.666

McNair, S., & Feeney, A. (2014a). When does information about causal structure improve statistical reasoning? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 67,* 625–645. http://dx.doi.org/10.1080/17470218.2013.821709

McNair, S., & Feeney, A. (2014b). Whose statistical reasoning is facilitated by a causal structure intervention? [Advance online publication]. *Psychonomic Bulletin & Review*. Advance online publication. http://dx.doi.org/10.3758/s13423-014-0645-y

Milligan, G. W. (1980). Factors that affect Type I and Type II error rates in the analysis of multidimensional contingency tables. *Psychological Bulletin, 87,* 238–244. http://dx.doi.org/10.1037/0033-2909.87.2.238

Morey, R. D., & Rouder, J. N. (2013). *BayesFactor: Computation of Bayes factors for simple designs*. [Online computer software manual]. Retrieved from http://CRAN.R-project.org/package=BayesFactor

Newell, B. R. (2013). Judgment under uncertainty. In D. Reisberg (Ed.), *Oxford Handbook of Cognitive Psychology* (pp. 602–615). New York, NY: Oxford University Press.

Newell, B. R., Koehler, D. J., James, G., Rakow, T., & van Ravenzwaaij, D. (2013). Probability matching in risky choice: The interplay of feedback and strategy availability. *Memory & Cognition, 41,* 329–338. http://dx.doi.org/10.3758/s13421-012-0268-3

Newell, B. R., & Rakow, T. (2007). The role of experience in decisions from description. *Psychonomic Bulletin & Review, 14,* 1133–1139. http://dx.doi.org/10.3758/BF03193102

Rakow, T. R., Demes, K., & Newell, B. R. (2008). Biased samples not mode of presentation: Reexamining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes, 106,* 168–179. http://dx.doi.org/10.1016/j.obhdp.2008.02.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. http://dx.doi.org/10.3758/PBR.16.2.225

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical applications*. Mahwah, NJ: Erlbaum.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804. http://dx.doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology, 60,* 158–189. http://dx.doi.org/10.1016/j.cogpsych.2009.12.001

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Frequency Processing and Cognition* (pp. 21–36). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780198508632.003.0002

Received August 25, 2014
Revision received October 13, 2014
Accepted November 7, 2014 ■