On the generality and cognitive basis of base-rate neglect

Elina Stengård[a*], Peter Juslin[a], Ulrike Hahn[b], Ronald van den Berg[a,c]

[a]Department of Psychology, Uppsala University, Uppsala, Sweden
[b]Department of Psychological Sciences, Birkbeck University of London, London, UK
[c]Department of Psychology, Stockholm University, Stockholm, Sweden

*Corresponding author
E-mail: elina.stengard@psyk.uu.se

ABSTRACT - Base rate neglect refers to people's apparent tendency to underweight or even ignore base rate information when estimating posterior probabilities for events, such as the probability that a person with a positive cancer-test outcome actually does have cancer. While many studies have replicated the effect, there has been little variation in the structure of the reasoning problems used in those studies. In particular, most experiments have used extremely low base rates, high hit rates, and low false alarm rates. As a result, it is unclear whether the effect is a general phenomenon in human probabilistic reasoning or an anomaly that applies only to a small subset of reasoning problems. Moreover, previous studies have focused on describing empirical patterns of the effect and not so much on the underlying strategies. Here, we address these limitations by testing participants on a broader problem space and modelling their response at a single-participant level. We find that the empirical patterns that have served as evidence for base-rate neglect generalize to the larger problem space. At the level of individuals, we find evidence for large variability in how sensitive participants are to base rates, but with two distinct groups: those who largely ignore base rates and those who almost perfectly account for it. This heterogeneity is reflected in the cognitive modeling results, which reveal that there is not a single strategy that best captures the data for all participants. The overall best model is a variant of the Bayesian model with too conservative priors, tightly followed by a linear-additive integration model. Surprisingly, we find very little evidence for earlier proposed heuristic models. Altogether, our results suggest that the effect known as "base-rate neglect" generalizes to a large set of reasoning problems, but may need a reinterpretation in terms of the underlying cognitive mechanisms.

Keywords: Base-rate neglect, Bayesian inference, Cognitive modeling

Introduction

Background

The question of whether the human mind adheres to the rules of probability theory has been debated ever since probability theory itself was developed a couple of hundred years ago. Since then the view has shifted drastically from Laplace's idea that probability theory is, in essence, common sense reduced to calculus (Brookes et al., 1953), to Kahneman's and Tversky's claim that people are unable to follow the rules of probability and instead have to rely on simple heuristics which often lead to fairly accurate answers but at other times produce large biases (Kahneman & Tversky, 1973). A more recent suggestion is to rather emphasize that the use of an adaptive toolbox of fast and frugal heuristics (Gigerenzer & Todd, 1999) takes positive advantage of information in real environment to make useful inferences, in practice, approximating close to normative behaviors by much simpler means.

A phenomenon that has inspired, and been used to exemplify, this research on heuristics and biases is *base-rate neglect*, people's tendency to respond to the evidence immediately at hand, while ignoring the base-rate (or prior) probability of an event. Although the original context of base-rate neglect emphasized that it is caused by people's reliance on simplifying heuristics (Kahneman & Tversky, 1973), the explanation and interpretation of the phenomenon remains debated to this day. One alternative explanation is that people are reliant on a toolbox of heuristics that are useful in the right context but will lead to biases, such as base-rate neglect, when the conditions are wrong (Gigerenzer & Hoffrage, 1995). Another approach emphasizes that people address base-rate problems much like any other multiple-cue judgment task (Brehmer, 1994; Karelia & Hogarth, 2008). On this view, people typically have a qualitative understanding that both the evidence and base-rate is relevant in base-rate problems, but they typically add up these cues, rather than spontaneously engaging in the multiplication prescribed by probability theory (Juslin, Nilsson, Winman, 2009). A third proposed account is that the underlying information integration is in fact consistent with probability theory, but corrupted by random noise in the process that appears as base-rate neglect in the empirical data (e.g., Costello & Watts, 2014). In this article, we will for the first time systematically compare these alternative accounts of base-rate neglect by computational modeling at the level of individual participants.

Base-rate neglect

One task that often has been used, both to argue for and against people's rationality, is the medical diagnosis task. Since the introduction of this task (Casscells et al., 1978) it has been formulated in many different ways. A typical example is the following:

> Suppose that 0.1% of all people in a population carry a virus. A diagnostic test for this virus detects it in 100% of the people who have the virus, but also gives "false alarms" on 5% of the people who do not have the virus. What is the chance that a person with a positive test result actually has the virus, assuming you know nothing about the person's symptoms or signs?

The correct answer can be computed using Bayes' theorem and gives a probability of ~2%. In the formulation above, the problem is specified in a "normalized" format, meaning that the information regarding the base rate (prevalence), hit rate (probability that a carrier of the virus is tested positive), and false alarm rate (probability that a non-carrier of the virus is tested positive) is given as percentages or single-event probabilities. People tend to overestimate the correct answer substantially often giving responses close to 100%. It has been argued that this is because of a tendency to respond 1 minus the false alarm rate (95% in this example) in such situations --possibly due to misinterpreting the false alarm rate as the total error rate (Tversky & Kahneman, 1981). If the hit rate is less than 100%, people often give answers that are close to the hit rate (McKenzie, 1994). While people thus seem to have different strategies depending on the value of the hit rate, they have in common that the diagnostic value of the hit rate and false alarm rate is overestimated while the base rate is ignored – a phenomenon commonly referred to as *base-rate neglect* or the *base rate fallacy* (Meehl & Rosen, 1955).

Evidence for base-rate neglect has been found in numerous studies, using different variations of Bayesian inference tasks (see Bar-Hillel, 1980; Barbey & Sloman, 2007; Kahneman & Tversky, 1973, for a few highly influential examples). The level of neglect varies, but the number of correct responses is seldom above 20% (see McDowell & Jacobs, 2017 for a meta-analysis of 35 studies). A complicating factor in explaining the effect is that its magnitude depends on the structure of the task. A number of facilitating factors has been explored that increase participants' use of base rates in Bayesian inference tasks, such as manipulating the base rate within subjects (Ajzen, 1977; Birnbaum & Mellers, 1983; Fischhoff et al., 1979), emphasizing the relevance of the base rate by highlighting a causal link to the inference task (Ajzen, 1977; Bar-Hillel, 1980; Fishbein, 2015), by providing explicit feedback, and by

considerable training (Goodie & Fantino, 1999). What these manipulations have in common is that they make decision makers more sensitive to base rates.

Although base-rate neglect is a well-established fallacy in the decision making literature by now, there are also studies that have found that people do respond to both the base rate and the hit rate, and instead are neglecting false alarm rates (Juslin et al., 2011). This relates to another classic judgment phenomenon called "pseudo-diagnosticity" (Ofir, 1988). This is the tendency of people to be influenced by diagnostically irrelevant information and/or disregarding actually informative information (but see also Crupi et al., 2009). In the medical diagnosis task, this could be manifested as being influenced by a high hit rate without taking into account a high false-alarm rate.

In the more recent debates, the focus has been on the format in which the information is presented. As originally shown in a study by Fiedler (Fiedler, 1988) and later further investigated by Gigerenzer and Hoffrage (1995), it appeared that more people are able to reason in accordance with Bayes' rule when all information is given in terms of *naturally sampled frequencies* (e.g. "95 out of 100 tested people" instead of "95%")[1]. The representation of the information has to be taken into account since algorithms, these authors claim, are information-format specific. According to their theory, the reason for people's poor performance on the medical diagnosis task and similar Bayesian inference tasks is that the cognitive algorithms are not adapted to compute with *normalized* frequencies, that is, probabilities or percentages. Although the problems are mathematically equivalent in the two formats, they are treated differently by the human mind, which, according to their theory, has evolved to reason with counts (natural frequencies) of events, but not with percentages or probabilities (normalized frequencies). The medical diagnosis problem above can be translated to a natural frequency format as follows:

> Suppose that one person in a population of 1,000 people carries a particular virus. A diagnostic test for this virus gives a positive test result on the person carrying the virus as well as for 50 of the 999 healthy persons. What is the chance that a person with a positive test result actually has the virus, assuming you know nothing about the person's symptoms or signs?

[1] For notational convenience, we will refer to this as "natural frequencies" throughout the rest of the paper.

By formulating the problem in terms of natural frequencies, Gigerenzer and Hoffrage found that the proportion of correct answers increased to approximately 50% of the cases compared to 16% with the normalized format. However, the reason why a natural frequency format is beneficial is still debated. Gigerenzer and Hoffrage's argument based on the evolution of the human mind has been contested by proponents of the nested-sets hypothesis (Barbey & Sloman, 2007; Sloman et al., 2003). They do not make any claims regarding the evolution of the human brain but instead base their argument on dual-process theory (Evans & Stanovich, 2013; Sloman, 1996). They argue that the effect of using frequencies is that it clarifies the probabilistic interpretation of the problem and makes the nested-set relations between the components evident. This in turn prompts people to shift from using a primitive associative judgment system to a deliberate rule-based system. The problem can therefore be formulated in either frequencies or probabilities but as long as the nested-set relations are transparent base-rate neglect will be reduced (Sloman et al., 2003).

In addition to using different numerical formats in Bayesian inference tasks, some studies have also used different visual formats to present the relevant information. In the two examples above, the information (base rate, hit rate, false alarm rate) was presented in a symbolic format. Some studies have examined if humans reason differently when this information is also presented in a pictorial format, for example by using Venn diagrams to represent normalized frequencies or collections of dots to represent counts. Adding a pictorial representation of the information has in some cases been shown to enhance participant's performance (Brase, 2009; Garcia-Retamero & Hoffrage, 2013). However, the results are mixed and it is clear that not all visual representations are helpful (Khan et al., 2015). Importantly, pictorial representations can also be used as a way of providing probability information to participants without giving them exact numbers (see, e.g., Harris, Corner, & Hahn, 2009; Harris, De Molière, Soh, & Hahn, 2017). The sense of uncertainty that this produces may make the experimental paradigm more representative for human reasoning in natural environments, where knowledge about base rates, hit rates, and false alarm rates is rarely exact (Juslin et al., 2009).

Explanations of base-rate neglect

Although the phenomenon of base-rate neglect has been known since at least the 1950s (Meehl & Rosen, 1955) the psychological explanation behind it is still a subject of discussion.

Representativeness. The first explanation for the base-rate neglect phenomena was put forth by Kahneman and Tversky, who suggested that it is caused by people relying on the representativeness heuristic (Kahneman & Tversky, 1973). They used a task in which

participants were presented with personality descriptions of people who were drawn from a population consisting of a known proportion of lawyers and engineers. Based on the personality descriptions the participants were to predict if the randomly drawn individual was a lawyer or an engineer. The result was that people in general disregarded the base rate proportions of lawyers and engineers and based their predictions only on the personality descriptions. Kahneman and Tversky's explanation was that people predict by representativeness: that they assess the representativeness (or similarity) of the personal description to the prototypical member of the professional categories (e.g., of a lawyer). The same argument can potentially be made to explain the base-rate neglect in the medical diagnosis task; the probability assessment would then be based on how representative a positive test outcome is for a diseased versus a healthy person. If it is considered more representative for a diseased person then the probability that the person has the virus is predicted to be high. Although there have been attempts to formulate the representativeness heuristic into a computational model (Bordalo et al., 2020; Dougherty et al., 1999; Juslin & Persson, 2002; Nilsson et al., 2008), its application to the sort of base-rate task considered here has not been examined. A problem with the representative heuristic is that it predicts that base rates are always ignored entirely. However, many experiments suggest that base-rate neglect is not an all-or-nothing phenomenon, but can differ in severity based on moderating factors, such as the format in which the problem is presented. Hence, although the representative heuristic can possibly account for the base-rate neglect in some tasks, there is no obvious mechanism that accounts for the moderating factors.

Heuristic toolbox. Gigerenzer and colleagues claim that as long as the information is presented in the appropriate natural frequency format people often make the appropriate computations and do not commit the base rate fallacy (Gigerenzer & Hoffrage, 1995). If, however, the information is presented in a normalized format people will rely on different non-Bayesian heuristics, such as reporting the hit rate or the difference between the hit rate and the false alarm rate, some of which will lead to base-rate neglect effects. The effects of moderating factors on base-rate neglect can, in principle, be accounted for by shifts between the different heuristics, although the exact characteristics of such a mechanism have never been specified.

Linear-additive integration. On this view, biases are due to people using strategies that are well adapted to the cognitive constraints and to the constraints of a noisy real-life environment. People have been shown to be inclined to combine information in a linear and additive manner (Juslin et al., 2008, 2011). In the context of the medical diagnosis task this means that, to varying degrees, participants have the basic understanding that the base rate and the hit rate (and to some extent the false alarm rate) are relevant to the posterior probability, but

left with unaided intuitive integration of this information they tend to engage in additive rather than multiplicative integration. Base-rate neglect arises when a participant assigns too little weight to the base rate in comparison to the optimal weight (i.e., the weight that produces the best linear-addition approximation of the Bayesian responses). The moderating factors on base-rate neglect are accounted for by people using various contextual cues to determine the weighting of the base-, hit-, and false alarm rate. In sequential belief revision tasks it has long been known that rather than relying on Bayesian integration people tend to average the "old" and the "new" data (Hogarth & Einhorn, 1992; Lopes, 1985; Shanteau, 1975).

Random noise in judgment. A different explanation of base-rate neglect starts off with the proposition that people, in principle, process information according to the rules of probability theory. However, it is assumed that the input and/or the output is subject to random noise that produces regression-like effects that may look as if base rates are being neglected. The effects of moderating factors can to some extent be accounted for by changes in the magnitude of the random noise. There is a long string of models that fall in this category, with the "Probability theory plus noise" model (Costello & Watts, 2014) being the most well-known. Recently, a reinterpretation of this model has been proposed in the form of a "Bayesian sampler" model (e.g., Sanborn & Chater, 2016; Zhu, Sanborn, & Chater, 2020).

A Skeptic Bayesian. Lastly, we introduce a variation to the view that people are in principle adhering to probability theory. The "Skeptic Bayesian" is a decision maker who integrates information according to Bayes' theorem, but takes the stated values with "a pinch of salt" that are dampened against the observer's own priors. This would be a rational strategy for an observer who believes that stated proportions are often more extreme than the true underlying parameters. For example, through experience, an observer may have come to appreciate that a stated proportion of .99 is more likely to err by being too high than by being too low and may, therefore, adjust it to a lower value. Moderating factors can presumably, at least to some extent, be accounted for by differences in the dampening priors that operate. This model also bears resemblances with a recently proposed variant of the Bayesian sampler model (Zhu et al., 2020). There are no previous empirical tests of this model but it is related to at least one model that have been used in previous studies. Birnbaum and Mellers (Birnbaum & Mellers, 1983) introduced a "subjective Bayesian model" in which objective probabilities are transformed to subjective probabilities in an at least conceptually similar manner as in our study. The exact implementation is however quite different from the Skeptic Bayesian; while the skeptic Bayesian has 7 free parameters, Birnbaum and Mellers' subjective Bayesian model has 22 free parameters.

Limitations in current literature

The medical diagnosis task and its many variants have been manipulated in several ways and used in numerous studies. While these studies have provided key insights into how humans reason with probabilities, they have also left many questions unanswered. Firstly, since many studies used only one trial, very little is known about individual's decision strategies. These strategies are best studied using cognitive modelling at the level of individuals, which, obviously, is not possible when having only a single datapoint. Furthermore, many of the previous studies have only examined problems similar to the ones used in the examples above, that is, problems with an extremely low base rate (often 1/1000), a hit rate close or equal to one, and a false alarm rate close to zero (e.g., Khan et al., 2015; Sloman et al., 2003). Although there are studies that tested participants on more than one trial (e.g., Fischhoff et al., 1979; Gigerenzer & Hoffrage, 1995; Juslin et al., 2011), none so far have performed a systematic exploration of the whole space of possible stimulus values. Consequently, it remains unknown how representative the results obtained in a rather extreme "corner" of this space are for human reasoning in general.

Other factors that have not always been taken into consideration are the various limitations inherent to biological information processing, such as neural noise (Faisal et al., 2008), the cost of neural computation (Lennie, 2003), and limits on the precision with which neural systems can approximate optimal solutions (Beck et al., 2012). These limitations will cause imperfections in the decision process and their collective effect can be thought of as a form of "decision noise". This resulting variability is not only believed to be a major source of errors in perception (e.g., Drugowitsch, Wyart, Devauchelle, & Koechlin, 2016; Stengard & Van Den Berg, 2019), but may potentially also explain, at least in part, many of the biases that have been found in cognition (Erev et al., 1994; Hilbert, 2012). Building on that research, Costello and Watts have proposed that people's probability judgments follow the laws of probability theory but are affected by random noise (Costello & Watts, 2014, 2016, 2018). If decision noise is *not* taken into account in the medical diagnosis task, any participant with noise in their decision process will be classified as non-Bayesian, even if the participant was using a Bayesian strategy. Gigerenzer and Hoffrage (Gigerenzer & Hoffrage, 1995) showed that `Bayesian' answers increased when the task was written in natural frequencies instead of probabilities. They interpreted this as evidence that the mind is adapted to reasoning with frequencies, but not with probabilities. However, an alternative explanation that cannot be

excluded unless the effects of noise are studied, is that the normalized format may lead to noisier computations of Bayes' rule rather than an abandonment of it.

Lastly, the method used in most previous studies has been to manipulate how the problem is presented and then see how the participant's performance changes. What is lacking is an exploration of what kind of strategies people use to solve the task in the different formats. This can be examined using formal model comparison methods, which have been rare in this field (but see, e.g., Birnbaum & Mellers, 1983; Juslin et al., 2011).

Purpose of the present study

In sum, the previous literature on human reasoning with conditional probabilities suffers from a number of limitations. First, studies have often only used a single trial. Second, the base rate, hit rate and false alarm rate used in that trial are often only given values from one extreme corner of the stimulus space where the base rate is very low. Third, decision noise was typically not taken into consideration. Fourth, the focus has been on describing the performance on this task and not on describing which strategies people are using to solve it. To address these issues, we used an approach where the participants performed several trials of the task allowing for a larger exploration of the stimulus space. We used the medical diagnosis task with both normalized format and natural frequency format and systematically explored a large part of the space of possible stimulus values. We also varied the presentation format by adding a condition with pictorial representations of the base rates, hit rates, and false alarm rates, which allowed for testing whether decision making strategies differ when the judgments are based on uncertain assessments rather than exactly stated numbers. Finally, to go beyond describing people's performance and examine the kind of strategy they are using we employed a cognitive modelling approach comparing five models that represent different hypothesized cognitive mechanisms.

Methods

Data sharing

The data is available at https://osf.io/3vkad/

Participants

Forty participants (31 female, 9 male; mean age 25.6 years, age span 20-45 years) were recruited from the student population at the Department of Psychology at Uppsala University.

These participants were distributed across the two conditions with pictorial stimuli (see below for further description) and were compensated with cinema tickets or gift vouchers with a value equivalent to approximately $10 per hour. The two conditions with symbolic stimuli were carried out using the crowd-sourcing service Amazon Mechanical Turk (MTurk), for which a total of 189 participants were recruited. The qualification requirement for participating in the study were a Human Intelligence Task (HIT) approval rate greater than 98% and Number of HITs approved greater than 5000. Seven of these participants were excluded, because they did not complete the whole experiment. All analyses were performed on the data from the remaining 182 participants (65 female, 115 male, 2 other; mean age 34.1 years, age span 19-70). Participants were compensated $5 for their participation for approximately 30 minutes of work.

Experimental task

On every trial, participants received information about the base rate (*BR*) of a fictitious virus in a fictitious hospital. They were also informed about the hit- and false-alarm rates of a medical test designed to detect the existence of the virus. The task was to estimate the probability that a randomly chosen person from the hospital who had received a positive test result actually had the virus. Participants provided their answer in percentages in all of the conditions[2]. In the symbolic tasks the response was given by typing the numbers on a keyboard and in the pictorial tasks by clicking on a number line. Within each participant, we factorially crossed five base rates (0.1, 0.3, 0.5, 0.7, 0.9), three hit rates (0.5, 0.7, 0.9), and three false alarm rates (0.1, 0.3, 0.5), resulting in a total of 45 trials. Note that the stimulus values were restricted to the "sensible" part of the stimulus space in which the hit rates were at least 50% and false alarm rates were at most 50%. Between participants, we factorially crossed two frequency formats ("natural frequency" and "normalized frequency") and two visual presentation formats ("symbolic" and "pictorial").

The two conditions with symbolic presentation format (Figure 1A-B) were similar to how information was presented to participants in most previous studies using the medical diagnosis task. In this presentation format, base rates, hit rates, and false alarm rates were all presented numerically. In the two conditions with the pictorial presentation format (Figure 1C-D), the information was represented by means of "probability matrices" similar to the ones used by Harris et al. (Harris et al., 2009, 2017). In these matrices, every single square represents a

[2] While they were allowed to provide decimal precision, all participants provided their responses as integer numbers.

person. The colour of the square signaled whether the person had the virus (red) or not (green) and the presence of a plus sign represented a positive result on the medical test. Each matrix consisted of 27 by 27 squares, which presumably was large enough to discourage participants from explicitly counting them. Stimuli were generated using the Psychophysics Toolbox (Brainard, 1997) for Matlab. In the tasks with normalized format, the participants were shown three matrices, separately representing the base rate, hit rate, and false alarm rate (Figure 1C). For the natural-frequency tasks, all the information could in principle be presented within a single matrix, but to increase visibility the hit rate and the false alarm rates were separated on the screen (Figure 1D).

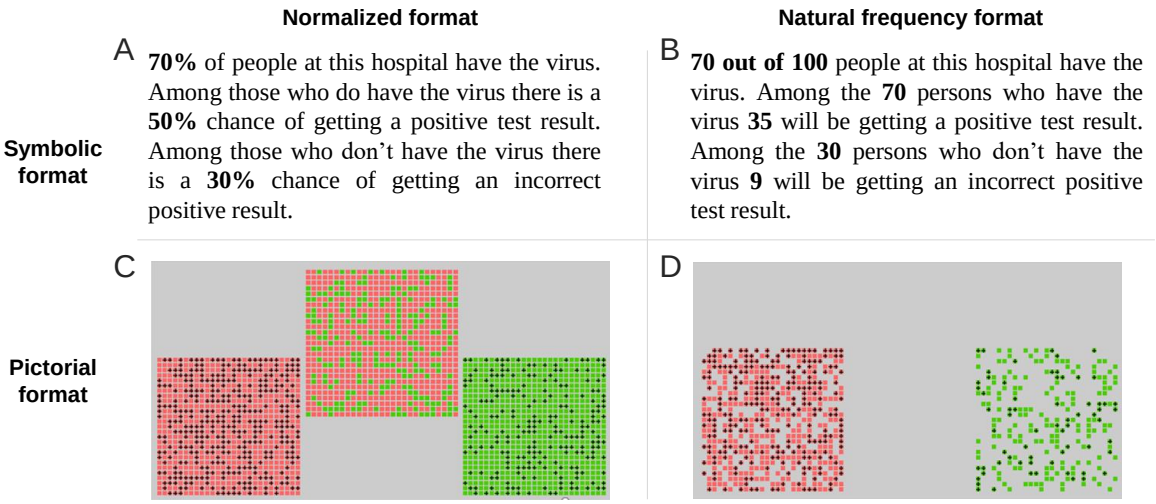

Figure 1 | Experimental design. Illustration of one trial in each of the four conditions. Top left: symbolic probability format. Top right: symbolic natural frequency format. Bottom left: pictorial probability format. Bottom right: pictorial natural frequency format. The pictorial examples are screenshots from the actual experiment, while the symbolic examples are translations of the original stimuli (which were presented in Swedish).

Experimental procedure

Symbolic task conditions. Data for these two conditions were collected online, using the Amazon Turk platform. The task consisted of 45 trials in which information was presented either in terms of natural frequencies (91 participants; Figure 1A) or as proportions (91 participants; Figure 1B). We randomized the trial sequence and then kept it the same for all participants. The starting point in this sequence varied between participants, in such a way that every test item was presented as the first trial for at least two participants. The participants in these conditions received general information about the experiment and gave informed consent before starting the task. They were encouraged to make their best judgments and performed the 45 trials in a self-paced manner. They were also informed that they were not allowed to use any

kind of calculator. Completion time varied strongly across participants, from 5 to 76 minutes (*Mdn* = 16 min).

Pictorial task conditions. The conditions with pictorial stimulus presentations were conducted in the lab. At the start of the session, the participant received general information about the experiment and gave informed consent. Thereafter, the experimenter left the room and the participant would start the experiment. Each participant performed the experiment with information presented to them either in natural frequency format (20 participants; Figure 1C) or in normalized format (20 participants; Figure 1D). The same 45 items as in the symbolic task were used, but now presented twice and with the trial order randomized per participant. Participants performed the experiment in a self-paced manner. Completion times varied from 6 to 69 minutes (*Mdn* = 17 min).

Discrimination task. After the main task, the participants in the pictorial conditions performed an additional discrimination task with 200 trials in which they were shown a single matrix (identical to the ones used in the main task) and were asked to estimate the proportion of squares that had plus signs on them. This task was used to assess the level of noise in the participant's estimations of the base rates, hit rates, and false alarm rates and was later used to put a constraint on the model parameters in the main task.

Control task. To verify that our paradigm is able to reproduce the original findings of base-rate neglect, we also collected data (N = 100) on a task with a single trial with the same stimulus values as the classical formulation of the task, base rate = 0.1%, hit rate = 100%, false alarm rate = 5% using symbolic presentation.

Models

To obtain insight into the strategies that participants used to solve this task, we fitted three types of process model to the experimental data: two Bayesian integration models, a linear-additive integration model, and two heuristics-based models. Each model implements a decision process that maps a triplet of observed input values (BR, HR, FAR) to a predicted response, R. This mapping consists of two stages (Figure 2). First, a deterministic integration rule is applied to map the input triplet to a decision variable d. Thereafter, a response R is generated by adding Gaussian "decision" noise to the log-odds representation of the decision variable (Zhang & Maloney, 2012). This noise is meant to capture imperfections in the decision strategies, such as those caused by neural noise and limitations in the amount of available neural resources. The standard deviation of this noise distribution, denoted σ, was a free parameter in all models. The

five tested models differed only with respect to the integration rule – the decision noise process was identical in all of them.

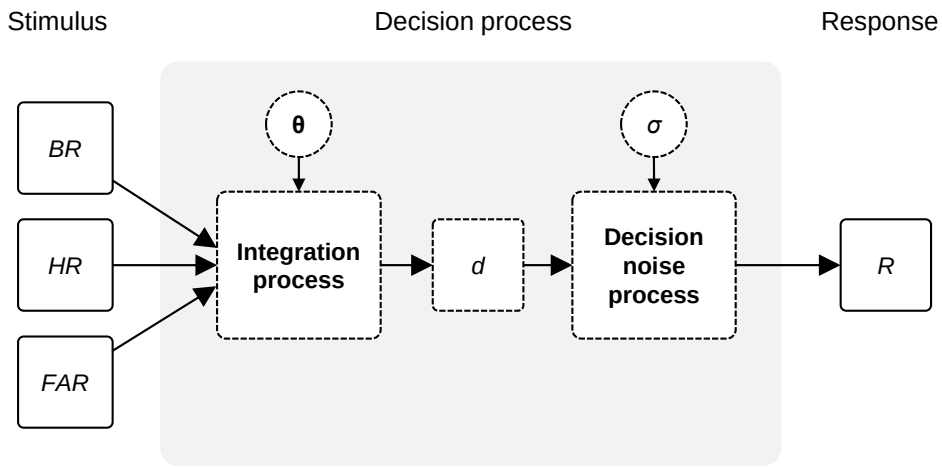

Figure 2 | Schematic overview of the process models considered in this study. On each trial, the model receives three inputs (*BR*, *HR*, *FAR*). These inputs are mapped to a decision variable, *d*, through a deterministic integration process that differs between models. Finally, the decision variable is mapped to a response, *R*, by corrupting it with Gaussian noise. Vector **θ** specifies the model parameters related to the integration process and *σ* is the standard deviation of the late noise distribution. Note that in the tasks with the pictorial input format, we assume that there is also noise on the stimulus inputs before they are integrated (see text for details).

Model 1: Imperfect Bayesian. This model originates in a long tradition of normative theories arguing that human cognition is based on Bayesian inference strategies (Griffiths & Tenenbaum, 2006; Oaksford & Chater, 1994; Tenenbaum et al., 2011). The Bayesian strategy for the experimental task is to calculate the posterior probability estimate via Bayes' rule. The response of this model is computed as

$$R = \frac{BR \times HR}{BR \times HR + (1 - BR) \times FAR} + \varepsilon \,,$$

where the first term on the right-hand side is the decision variable and the second term, ε, is the decision noise. This "imperfect" Bayesian model has in common with standard Bayesian models that it assumes that observers integrate information in accordance with Bayesian decision theory. However, it differs from those models in the sense that it allows for imperfections in the execution of this integration, due to factors such as neural noise and limits in the amount of available cognitive resources. We found in previous work that this type of

model captures human behavior (on a visual search task) better than the standard Bayesian model and all tested suboptimal heuristic models (Stengard & Van Den Berg, 2019).

This model shares similarities with the "Probability theory plus noise" model by Costello and Watts (Costello & Watts, 2014, 2016, 2018). The main difference is that the model by Costello and Watts puts the noise on the input representations (before integration), while we put it on the decision variable (after integration)[3]. Although it has been repeatedly proposed in the literature that base-rate neglect could be a side effect of random noise (Costello & Watts, 2018; Erev et al., 1994; Hilbert, 2012; Juslin et al., 1997), the hypothesis has rarely been subjected to empirical tests in experiments with cognitive modeling.

Model 2: Skeptic Bayesian. In addition to the Imperfect Bayesian model, we fitted a version that we call the "Skeptic Bayesian". This model represents a Bayesian observer with prior beliefs about base rates, hit rates, and false alarm rates in the context of viruses and medical tests. Instead of fully trusting the information provided by the experimenter, the skeptic Bayesian combines that information with its own priors before applying the integration rule. We modelled this skepticism by introducing Beta distributed priors on the base rate, hit rate, and false alarm rate, which are specified by six free parameters: α_{BR} and β_{BR} for the prior belief distribution over base rates; α_{HR} and β_{HR} for the prior belief distribution over hit rates; and α_{FAR} and β_{FAR} for the prior belief distribution over false alarm rates. These parameters can be interpreted as counts of prior outcomes that the observer has witnessed. For example, α_{BR}=20 and β_{BR}=180 would represent an observer who has witnessed 20 persons that did have a virus and 180 that did not. When this observer is told in the experiment that "*25 out of 100 people in this hospital have the virus*", she combines this information with her prior, which in this example would give a believed base rate equal to (20+25)/(200+100)=0.15. The model assumes that the information provided on each trial is based on a sample of 100 observations. Hence, if on a trial in the "normalized format" variant of the task it is stated that "Among those who do have the virus, there is a 90% chance of getting a positive test result", the model will use 90 and 10 as counts for the hit rate.

Model 3: Linear additive integration. The third model is based on findings from research on multiple-cue judgment tasks (e.g. estimate the price of an apartment based on its size, number of rooms, and floor) (Brehmer, 1974; Juslin et al., 2008). This research suggests that people tend to combine cues by linearly weighting and adding them. Since the task in the present

[3] We also tested a variant of the imperfect Bayesian model with noise on the input variables and found that they are indistinguishable in terms of goodness of fit.

study can be seen as such a task (with BR, HR, and FAR as cues), it is conceivable that participants used this strategy. Under a linear-additive integration rule, the process model takes the form

$$R = w_{\mathrm{BR}} BR + w_{\mathrm{HR}} HR + w_{\mathrm{FAR}} FAR + \varepsilon ,$$

where the weights w_{BR}, w_{HR} and w_{FAR} determine how much each piece of information contributes to the estimate of the posterior. The weights are fitted as free parameters with an unconstrained range. The linear additive account of base-rate neglect is both indirectly supported by the literature on multiple-cue judgment (Brehmer, 1994) and directly supported by the results from computational modeling on base-rate problems (Juslin et al., 2011).

Model 4: Heuristic toolbox The fourth model that we test is based on four heuristics that Gigerenzer and Hoffrage (Gigerenzer & Hoffrage, 1995) derived from self-reported strategies of a large number of participants performing a task similar to the one we use here. The first of these is the "joint occurrence" heuristic, which approximates the posterior as the product of the base rate and the hit rate, $d=BR\times HR$. This will generally underestimate the true posterior, but can serve as a decent approximation when BR is high, which according to Gigerenzer and Hoffrage is the kind of situation in which people use this heuristic. The second heuristic entirely ignores *BR* and *FAR* and simply approximates the posterior as the *HR*, $d=HR$. This "Fisherian" heuristic leads to the same result as Bayes' theorem when the base rate of the virus is equal to the base rate of positive test outcomes, that is, when $p(v) = p(t)$. Therefore, Gigerenzer and Hoffrage argue that people use this heuristic more frequently when the difference between $p(v)$ and $p(t)$ is small. The final two heuristics – referred to as "Likelihood subtraction" – are variants of an algorithm that ignores the base rate and seems to have been the predominant choice by participants in previous studies (Cosmides & Tooby, 1996). The first variant takes the difference between the hit rate and the false alarm rate, $d=HR-FAR$. The second variant is a simplification of this rule, in which the hit rate is assumed to be equal to 1, such that $d=1-FAR$.

A key difference between these and the previous models is that it adjusts the integration rule to the situation. This adds an extra step to the general model shown in Figure 2, namely that of choosing which integration rule (heuristic) to use on any given trial. While there are many different ways to do this, a general assumption underlying heuristic decision-making models is that people use heuristics to approximate correct answers without performing Bayesian or other complex computations. Therefore, it seems reasonable to assume that the

most likely heuristic used by a participant is the one that best approximates the correct answer on that trial. Under this assumption, the process model takes the form

$$R = \left(\underset{d_i^{\text{heuristic}}}{\operatorname{argmin}} \left| d_i^{\text{heuristic}} - d_{\text{Bayes}} \right| \right) + \varepsilon \,,$$

where $d_i^{\text{heuristic}}$, $1 \le i \le 4$, refers to the decision variables related to the four heuristic rules and d_{Bayes} is the optimal (correct) decision variable and $|\cdot|$ is the absolute value operator. The argmin operator selects the decision variable that most closely approximates the correct (Bayesian) decision variable.

Model 5: Lexicographic model. The final model that we tested is a "lexicographic" variant of the Heuristics Toolbox model. It uses the same four decision heuristics, but with a different way of choosing which one to use on each trial. Instead of choosing the objectively best heuristic for the given situation (Model #4), the lexicographic model assumes that participants consider the informativeness of each cue (*HR*, *BR*, *FAR*) in turn until a sufficiently informative cue is found. The order in which the model considers the cues is informed by the general pattern in the literature, which has shown that people respond strongly to hit rates, to some extent to the base rates, but rarely to the false-alarm rates. Therefore, the model first checks the value of the *HR*. If it is considered to be sufficiently informative about whether the event will happen (i.e., if it deviates sufficiently from .5), the participant simply reports the *HR*, in accordance with the Fisherian heuristic introduced above. If the *HR* is not sufficiently informative, it is assumed that the participant next considers whether the *BR* is informative about whether the event will happen (i.e., is far from .5) and reports it if it is. If neither the *HR* nor the *BR* is highly informative by itself and also the FAR is uninformative (i.e., close to .5), the participant reports *HR*×*BR*, as according to the Joint Occurrence heuristic introduced above. Finally, if FAR is very informative (i.e., far from .5), the participant integrates this additional information and reports the HR – FAR, as according to the Likelihood Subtraction heuristic above. We define a cue to be informative when it is outside the 0.50 ± δ range, with δ fitted as a free parameter. Formally, this model is formulated as

$$R = \begin{cases} HR + \varepsilon & \text{if} \quad HR \notin 0.50 \pm \delta \\ BR + \varepsilon & \text{if} \quad HR \in 0.50 \pm \delta \;\; \& \;\; BR \notin 0.50 \pm \delta \\ BR \times HR + \varepsilon & \text{if} \quad HR \in 0.50 \pm \delta \;\; \& \;\; BR \in 0.50 \pm \delta \;\; \& \;\; FAR \in 0.50 \pm \delta \\ HR - FAR + \varepsilon & \text{otherwise.} \end{cases}$$

Ecological rationality of the non-Bayesian models

To assess the "ecological rationality" of the three non-Bayesian models, we optimized their parameters with respect to minimizing the root mean squared error (*RMSE*) between their predicted responses and the correct (Bayesian) responses on the 45 trials in our experimental task (Figure 3). All three models perform reliably better than a randomly guessing observer and an observer who always responds 0.50, which indicates the models do something sensible. The optimized linear additive model consistently outperforms the two heuristic models and has a relatively small error compared to a randomly responding observer and an observer who always responds 0.50. The optimal weights in the linear-additive model are w_{BR}=1.11, w_{HR}=0.52, and w_{FAR}=−0.71, which means that the best performance is achieved by giving most weight to the base rate and least (in absolute terms) to the hit rate.

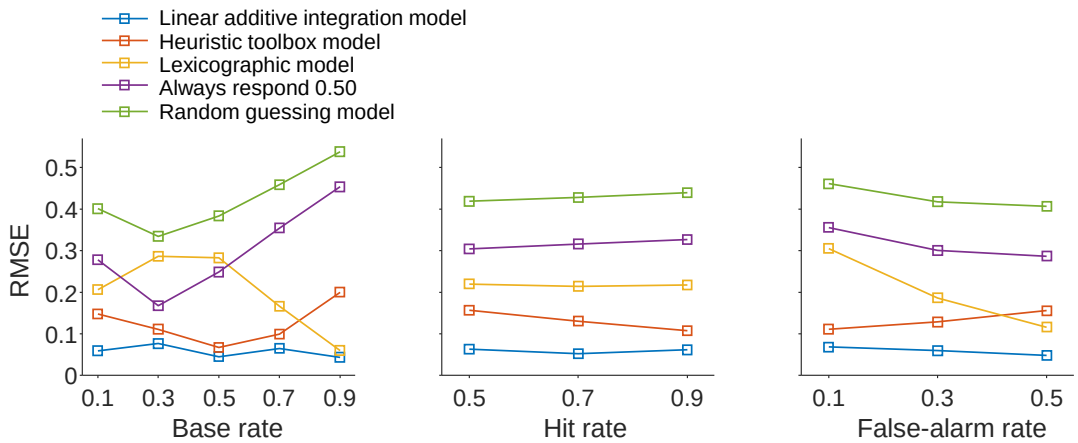

Figure 3 | Best possible performance of the non-Bayesian models on our experimental task. Accuracy of the three non-Bayesian models on our experimental task, after optimizing their parameters with respect to minimization of the root mean squared error (RMSE). All models perform better than a randomly guessing observer (green) and an observer who responds 0.50 on each trial, but with clear differences between them.

Parameter fitting and model comparison

We fitted the model parameters by using maximum-likelihood estimation, that is, by finding for each model the parameter vector $\boldsymbol{\theta}$ that maximized $\prod_{i=1}^{\#\text{trials}} p\left(D_i \mid \boldsymbol{\theta}, BR_i, HR_i, FAR_i\right)$, where D_i, BR_i, HR_i, FAR_i, are the participant's response and the presented base rate, hit rate, and false alarm rate on trial i, respectively. This maximization was done numerically, by using an

optimization algorithm based on Bayesian direct search (Acerbi & Ma, 2017). Likelihoods were computed using Inverse Binomial Sampling (IBS) (van Opheusden et al., 2020), which is a numerical method that samples responses from the model until it is equal to the participant's response. An advantage of this method is that – unlike many other numerical methods – it guarantees the likelihood estimates to be unbiased[4]. However, a disadvantage is that it can get stuck on parameter combinations that are unable to reproduce one or more of the participant's responses. To avoid this, we added a free lapse rate λ to each model. To avoid overfitting and to allow for a fair model comparison, the models were fitted using five-fold cross-validation. In each of the five runs, a unique subset of 20% of the trials was left out during parameter fitting. The log likelihood values of the left-out trials were summed across the five runs, providing a single "cross-validated log likelihood" value per model fit. To reduce effects of the optimization algorithm sometimes ending up in a local maximum, we performed each fit 20 times with a different starting point in each run.

Using the discrimination task to constrain the models in the pictorial conditions

In the pictorial conditions, participants estimated the base rate, hit rate, and false alarm rate from the presented "probability matrices" (Figs. 1C-D). An important difference with the symbolic condition is that these estimates were likely to be non-exact (or: "noisy"). Previous research suggests that people are equipped with a dedicated "approximate number system" (Dehaene, 1992) to make this kind of visual judgments. A well-known finding in that field of research is that the amount of noise on these estimates scales with the numerosity of the estimated set, which can be modelled using Weber's law (Shepard et al., 1975). Instead of modelling this noise as an additional free parameter in the models for the pictorial conditions, we estimated the value of this parameter from an independently performed discrimination task that was similar to typical tasks used in numerosity estimation experiments (see *Procedure* above). The value of the noise level parameter was estimated separately for each participant, by fitting a model that assumed that people's observations of numerosity are corrupted by Gaussian noise that scales with the magnitude of the numerosity.

[4]The presence of early noise in models fitted to data from the pictorial conditions prohibited us from computing likelihoods analytically.

Results

Effect of base rate on responses

Control experiment. The control experiment consisted of a single trial in which participants were presented with the classic formulation of the base-rate neglect task (base rate: 0.1%; hit rate: 100%; false alarm rate: 5%; correct answer: 1.96%). When counting all answers between 1.8% and 2.2% as correct[5] (Sloman et al., 2003), 9% of the participants were classified as giving the correct answer (Figure 4A). This is consistent with a meta-analysis of 115 previously reported experiments, where the majority of the observed proportions of correct answers on , his task was below 20% (McDowell & Jacobs, 2017). The modal response in our control task was "95%" (~20% of the responses), which is also consistent with previous findings and which is often interpreted as base-rate neglect (Sloman et al., 2003). Hence, the control experiment successfully replicates earlier findings of base rate neglect. Note, however, that while the *modal* response was 95%, a majority of the participants gave an answer that was close to correct, thus not showing evidence for base rate neglect.

Main experiment. The task in the main experiment was the same as in the control experiment, except that it consisted of a much larger number of trials and variation in the base rates, hit rates, and false alarm rates. Accuracy levels differed slightly between the four conditions of the main task (Figure 4B-C). Bayesian *t*-tests confirmed that the average performance was reliably better than that of a participant making random guesses on every trial, which would produce a mean absolute error of 35 ($BF_{10} > 1.9 \cdot 10^{17}$ in all four conditions). A two-way Bayesian ANOVA with mean absolute error as the dependent variable and task format and presentation format as independent variables revealed strong evidence for an effect of task format ($BF_{inclusion} = 23.19$), consistent with previous reports (Gigerenzer & Hoffrage, 1995) that people are more accurate when the information is presented in terms of natural frequencies ($M = 18.02$, $SD = 12.34$) compared to normalized formats ($M = 23.20$, $SD = 10.20$). Moreover, the same test showed anecdotal evidence against both an effect of presentation format ($BF_{inclusion} = 0.46$) and an interaction effect ($BF_{inclusion} = 0.37$). Hence, somewhat unexpectedly, performance was comparable between the pictorial and symbolic conditions and the advantage of presenting the information in terms of natural frequencies was also comparable between these two conditions.

[5] All participants responded with an integer, meaning that in practice only answers of "2%" were considered correct.

One of the main questions of this study is whether the base rate neglect effect generalizes to inference problems different from the classic one with a high hit rate, a low false alarm rate, and an extremely low base rate. If all our participants used a decision strategy that ignored the base rate, then we should find that their average responses are similar across sets of trials with the same hit- and false-alarm rate values. Our data show that this is clearly not the case (Figure 4C right). In all four conditions, participants on average increased their responses in reaction to an increase in the base rate and hit rate and decreased their response in reaction to an increase in the false alarm rate (three-way repeated measures ANOVA[6]; $BF_{inclusion}$ was so large it exceeded the numerical precision of JASP for all main effects). Although the participants were sensitive to the information, they did not adjust their answers enough from a normative perspective. To examine this in more detail, we quantified each observer's sensitivity to the base rate, S_{BR}, as the slope of a linear fit to their responses divided by the slope of a linear fit to the Bayesian observer's responses. A participant who entirely ignored the base rate will have a sensitivity of 0, while a participant who reacted equally strongly to changes in the base rate as the Bayesian observer has a sensitivity of 1 (Figure 4E). Note that "reacting equally strongly as the Bayesian observer" does not mean to perform perfectly in accordance with Bayes' theorem. It is possible to produce incorrect responses while still adjusting the responses correctly to accommodate a change in the base rate. See the right plot in Figure 4E for an example of such a participant. We found that the sensitivity values are largely clustered around 0 and 1 (Figure 4F), which suggests that many of the participants either entirely ignored the base rate or entirely accounted for it, respectively. This finding is qualitatively consistent with the data obtained using the classic base-rate neglect task, where we see a similar kind of clustering (Figure 4A).

These results give rise to two preliminary conclusions. First, there are large individual differences in how participants behave in both the classic formulation of the task and our generalization of it. This warrants caution when making population-level statements about base rate neglect. Second, the base-rate neglect effect does not seem to be limited to inference problems with a high hit rate and extremely low base rate: in both the control task and the generalized task, participants largely cluster into a group that almost entirely accounts for base-rate neglect and another group that almost entirely neglects base rates.

[6] The different values of the base rate, hit rate and false alarm rate were entered as repeated measures factors and individual subject responses as dependent variable.

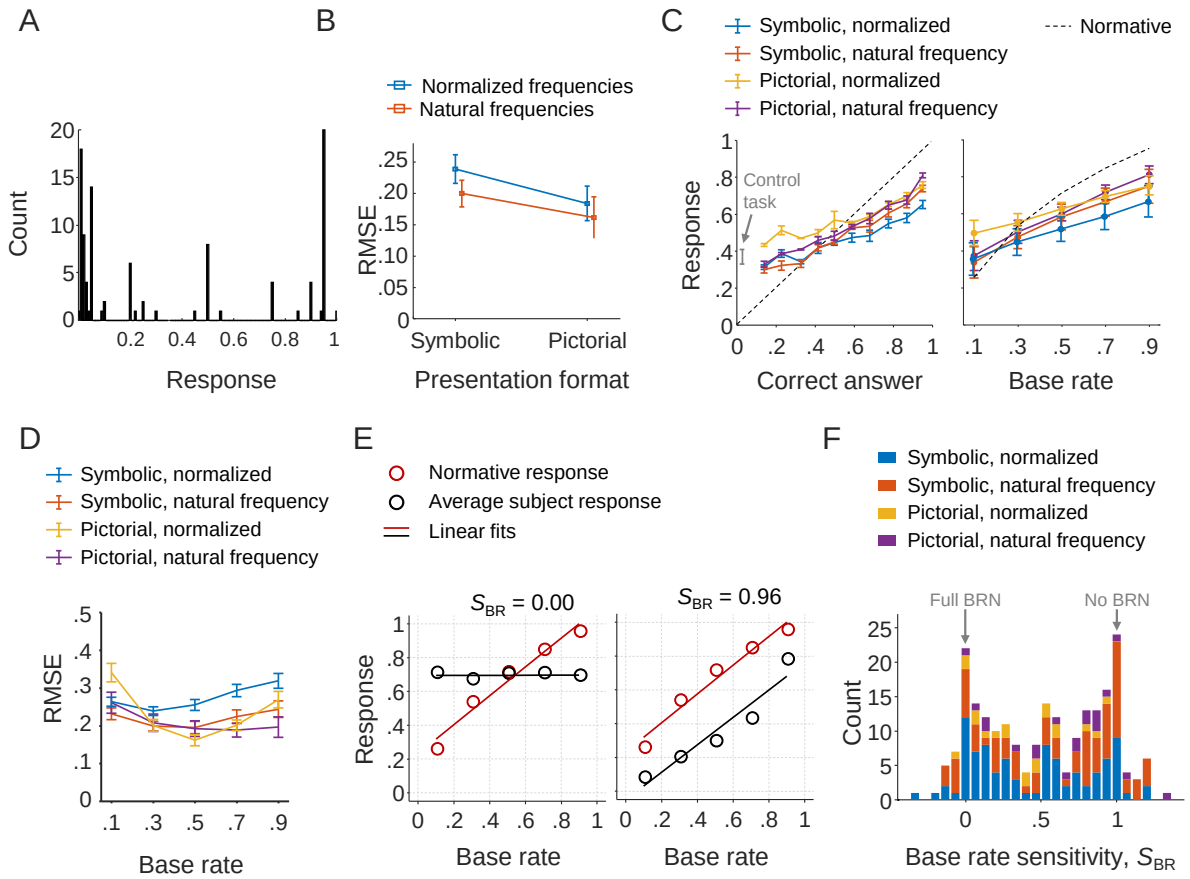

Figure 4 | Analysis of participants' sensitivity to base rates. (A) Distribution of responses in the control experiment. (B) Subject-averaged response accuracy in the main experiment, split by condition. RMSE = Root Mean Squared Error. (C) Left: Subject-averaged responses binned by the correct response and split by condition. Right: Subject-averaged responses for each base rate, collapsed across hit rates and false alarm rates and split by condition. See Appendix for corresponding plots of hit rate and false alarm rate. (D) Subject-averaged response accuracy for each base rate, collapsed across hit rates and false alarm rates and split by condition. (E) Left: an example of a participant whose average response was independent of the base rate; the base rate sensitivity, S_{BR}, was computed as the ratio between the linear fit slopes. Right: an example of a participant who responded to the base rate almost equally strongly as the normative observer. (F) Distribution of sensitivity values across all subjects. BRN = Base Rate Neglect.

Model comparison

We see two plausible explanations for the individual differences identified above. A first possibility is that all participants used the same decision-making strategy, but with different cognitive parameters. For example, they may all have been solving the problems using linear-additive integration, but with individual differences in the weights assigned to the *BR*, *HR*, and *FAR*. Alternatively, it can be that different participants used categorically different decision-making strategies. For example, some of them may have used a Bayesian strategy while others used a heuristic strategy. To get insight into the decision strategies used by the participants, we

fitted five cognitive models to their individual data sets: two variants of the Bayesian model, a linear-additive integration model, and two heuristic integration models (see *Task and Models* above for details).

Symbolic stimuli. When comparing the models based on their cross-validated log likelihood values, the Skeptic Bayesian was selected for more participants than any other model and had the highest group-averaged log likelihood (Figures 5A and 6A). However, just as in the earlier analysis, we found large individual differences in model preference. In fact, despite the Skeptic Bayesian being the overall best model, it was the preferred model for less than half of the participants (71 out of 182, i.e., 39.0%). The linear-additive model was the runner-up model, being the selected one for a total of 56 participants (30.8%). The imperfect Bayesian was the preferred model for 33 participants (18.1%), the lexicographic model for 21 participants (11.5%), and the heuristic toolbox model for only 1 participant (0.6%). Hence, from the perspective of model classes, the evidence was strongest for the Bayesian integration models (preferred in 104 participants, 57.1%), weaker for the linear-additive integration model (56 participants, 30.8%), and overall weakest for the heuristic models (22 participants, 12.1%).

Altogether, the modelling results looked very similar between the normalized-frequency condition and the natural-frequency condition. However, a noticeable difference is that the Imperfect Bayesian was the preferred model in the condition with the natural frequency format more than twice as often than in the condition with the normalized format (23 *vs.* 11 times). This suggests that more participants acted like a Bayesian when information was presented in the natural frequency format, which is consistent with earlier literature as well as with the accuracy difference reported above (Figure 4B).

All models provided decent accounts of the group-averaged responses as a function of the base rate, but with the Skeptic Bayesian and Linear Additive models again doing visibly better than the other models (Figures 5B and 6B). This difference was also apparent at the level of individual trials, where the responses of the Skeptic Bayesian model and the Linear Additive model were more strongly correlated with the participant responses than the responses of the other models (Figures 5C and 6C). The explained variance (based on the best-fitting model for every participant) varied substantially between participants with the median being .66 and .42 for the natural frequency and normalized frequency formats respectively. These values are most likely underestimated due to the level of random noise in the data as we will see for the pictorial conditions below.

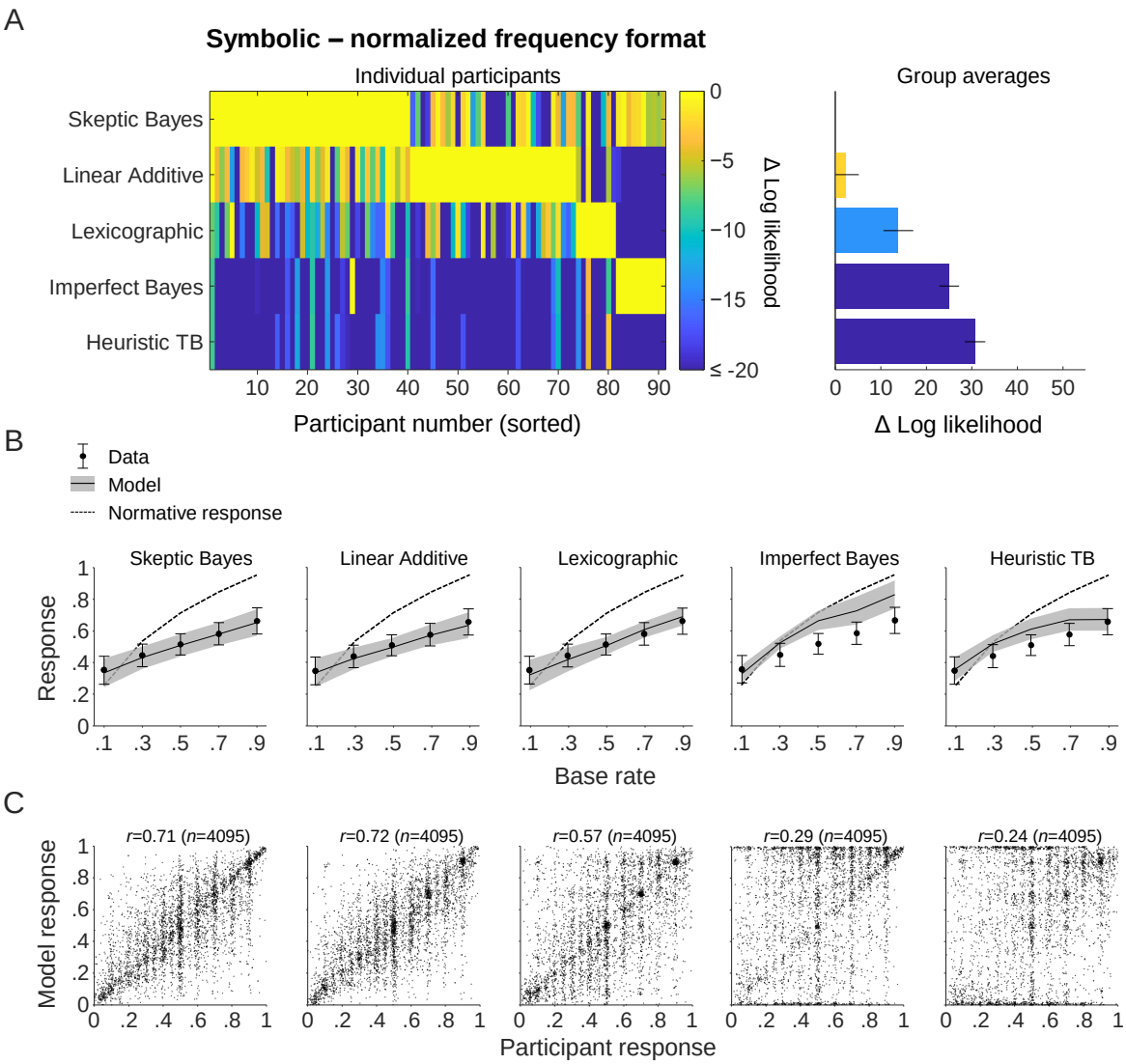

Figure 5 | Modeling results for the condition with symbolic stimuli presented in probability format. (A) Left: Cross-validated model log likelihoods relative to the best model. The preferred model for each participant is indicated in yellow and the worst models are indicated in blue. For visualization purposes, participants were sorted in such a way that all participants on which a particular model was the preferred one would line up (yellow areas). Right: Cross-validated model log likelihoods relative to the Skeptic Bayesian model, averaged across all participants. Errors bars indicate 1 s.e.m. (B) Participant-averaged responses as a function of the base rate. Error and shaded areas indicate 1 s.e.m. (C) Model responses under maximum-likelihood parameters plotted against participants' responses. The panels are in the same order as in 5B. For visualization purposes, each data point was jittered in both directions by Gaussian noise with a standard deviation of 0.01; *r* indicates the Pearson correlation coefficient and *n* the number of responses.

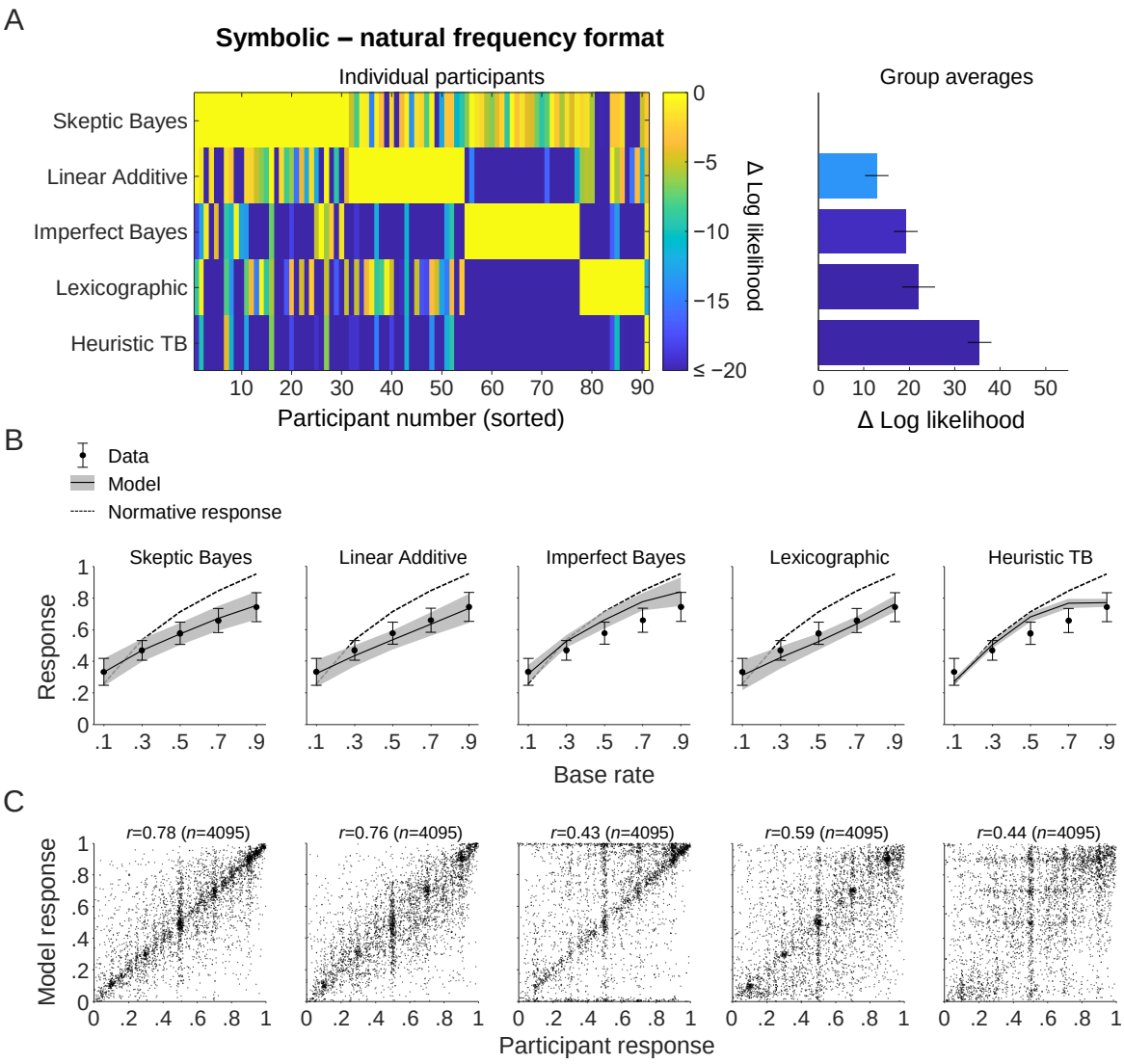

Figure 6 | Modeling results for the condition with symbolic stimuli presented in natural frequency format. This figure follows the same layout as Figure 5.

Pictorial stimuli. The modeling results for the pictorial conditions were very similar to those for the symbolic conditions. In both the normalized format and natural frequency format conditions, the Skeptic Bayesian model was again favored for more participants than any other model, and again with the Linear Additive model as the runner-up (Figures 7A and 8A). Moreover, these two models accounted well for the group-level responses as a function of base rate, with the other models doing visibly worse (Figures 7B and 8B). The most striking difference is that the Skeptic Bayesian is now favored more forcefully, in the sense that it is selected for 75.0% of the participants (30 out of 40), compared to the 39.0% in the symbolic conditions. Another difference is that the trial-level correlations between participant and model responses (Figures 7C and 8C) were lower than in the symbolic conditions. This was expected

because of the presence of early noise in the pictorial conditions. Since every trial was performed twice in this condition we can compute the true variance in the data and use this to adjust the estimates of the explained variance by dividing the explained variance with the true variance. The median adjusted explained variance then become .86 and .74 for the natural frequency and the normalized frequency format respectively.

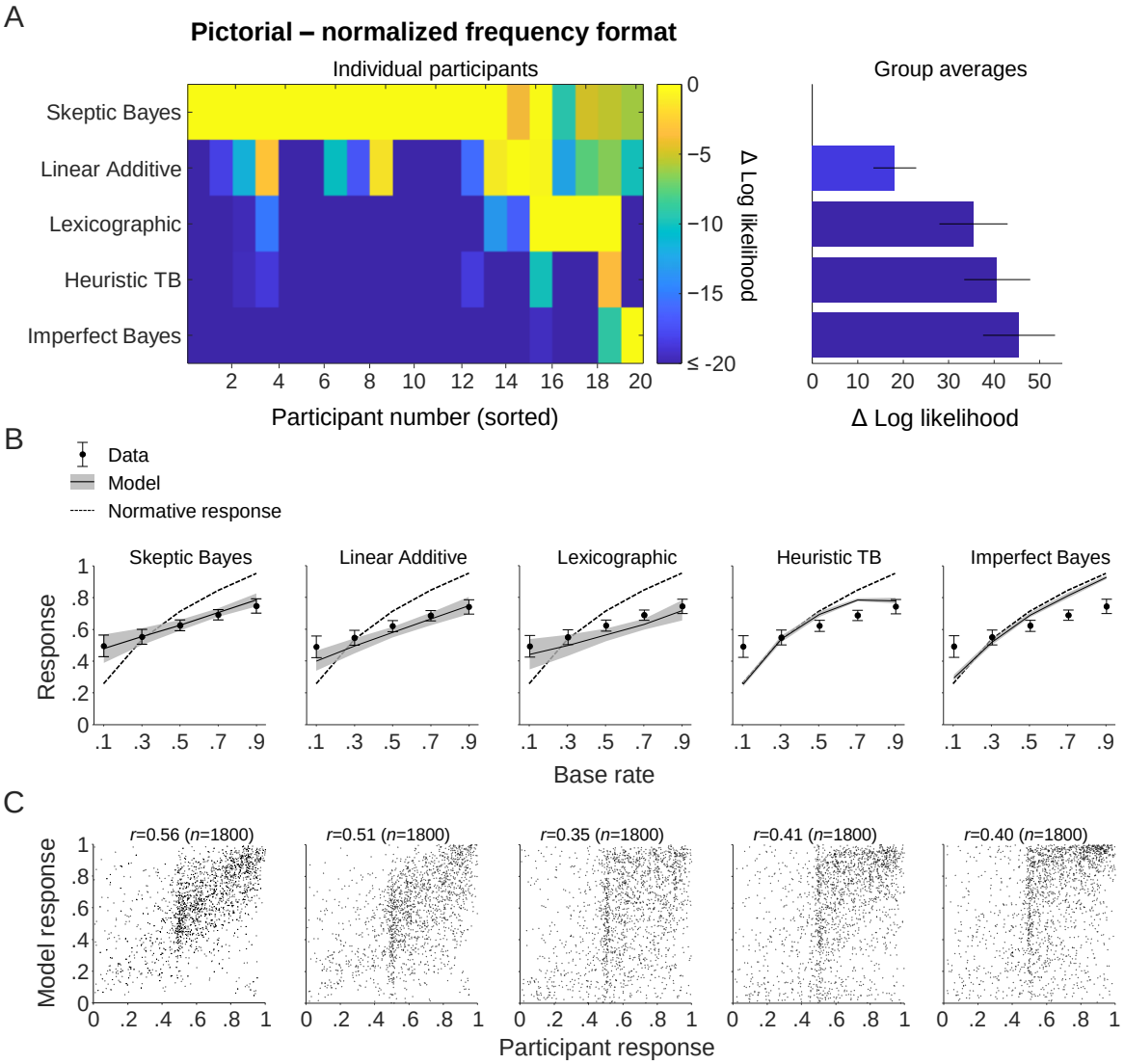

Figure 7 | Modeling results for the condition with pictorial stimuli presented in probability format. This figure follows the same layout as Figure 5.

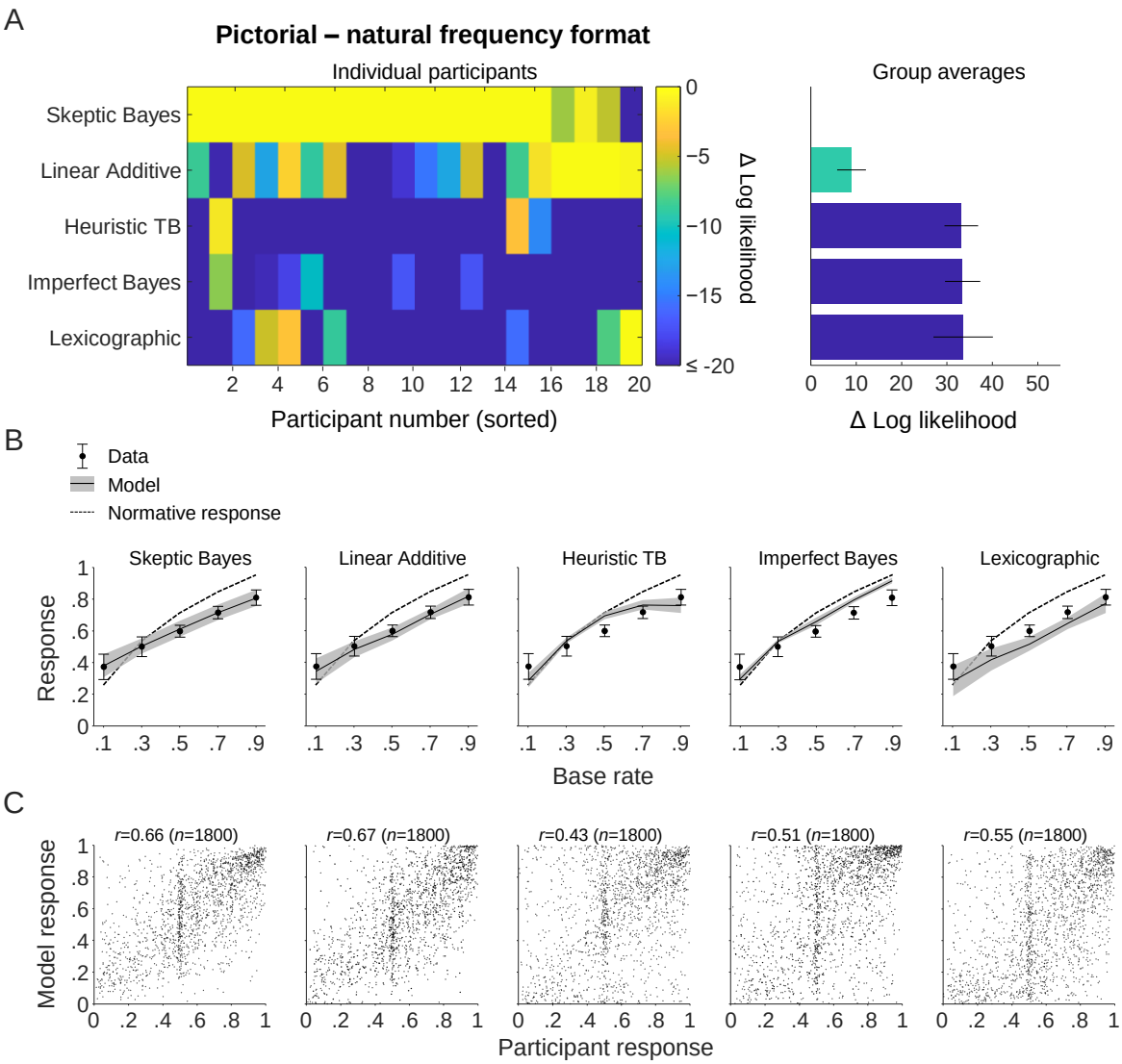

Figure 8 | Modeling results for the condition with pictorial stimuli presented in natural frequency format. This figure follows the same layout as Figure 5.

Parameter estimates. To further investigate how participants made use of the base rate, hit rate and false alarm rate in their judgments, we next looked at the estimated weights from the two best fitting models: The Skeptic Bayesian model and the Linear Additive integration model. In the Linear-Additive model, we found that too little weight was given to both the base rate and the false-alarm rate in all four conditions (Figure 9). On average, the hit rate was weighted correctly, but with large individual differences in the estimated weights. In the Skeptic Bayesian model, we found large individual differences in the estimated priors (Figure 10). Hence, from the perspective of this model, our participants had largely varying beliefs about base rates, hit rates, and false alarm rates.

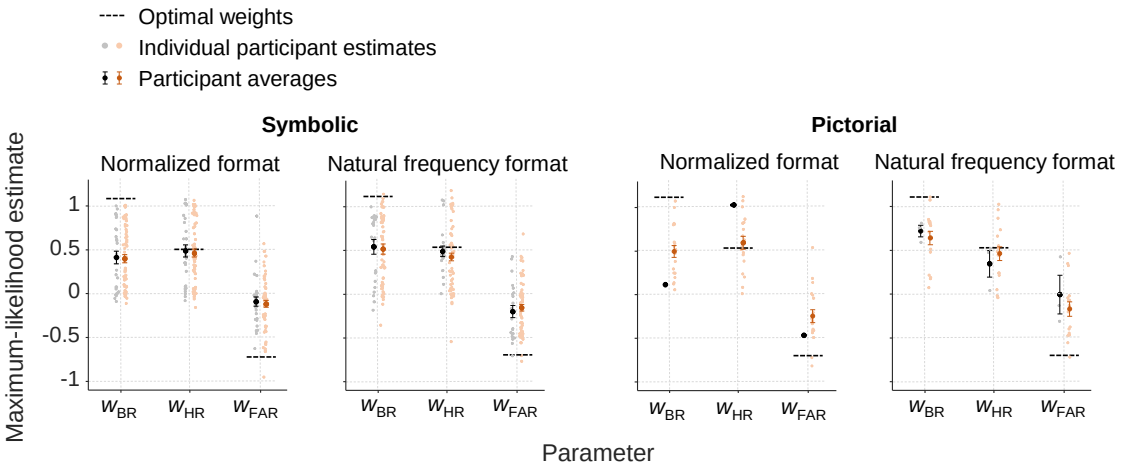

Figure 9 | Estimated weights of the linear additive model. Estimated weights for the base rate, hit rate and false-alarm rate for the linear additive model. Black error bars and dots indicate the mean and individual weights of participants for whom the linear additive model provided the best fit. Red error bars and dots indicate the mean and individual weights of all other participants. The dashed lines indicate the weights of an additive integration model fitted to responses from a Bayesian decision maker (i.e., the linear additive weights that best approximates the Bayesian solution).

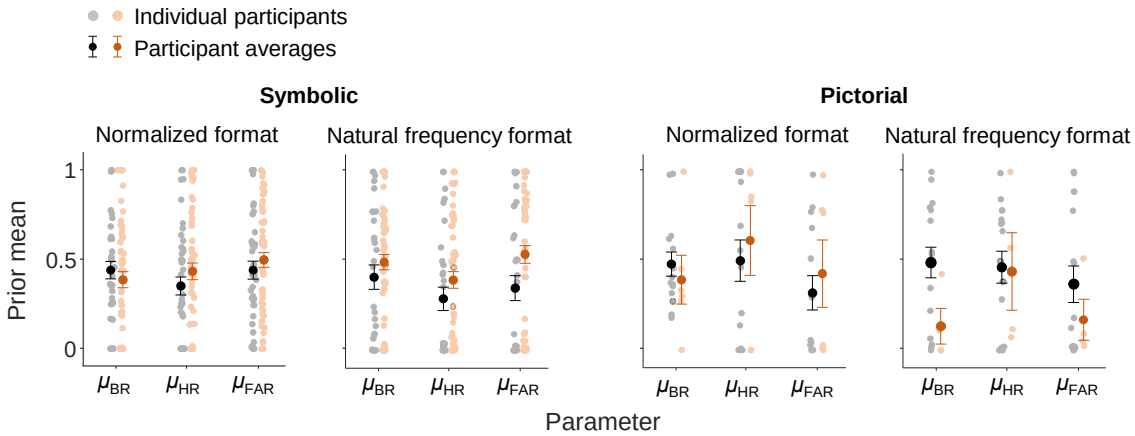

Figure 10 | Estimates of base rate priors in the Skeptic Bayesian model. Black error bars and dots indicate the mean and individual weights of participants for whom the linear additive model provided the best fit. Red error bars and dots indicate the mean and individual weights of all other participants.

Analysis of the relation between decision strategies and base rate sensitivity

The results so far suggest that there are large individual differences both in the extent to which participants accounted for base rates in their decisions (Figure 4F) and in the preferred cognitive model fitted to their data. We next examined whether these two findings are related: is there

any evidence in the modeling results to suggest that participants with little base rate neglect were using a categorically different strategy than participants with strong base rate neglect? Or, alternatively, is the strength of base-rate neglect mainly a matter of differences in cognitive parameter values? To answer this, we replotted the distribution of sensitivity values shown in Figure 4F, but now color-coded by the best-fitting model for each participant (Figure 11A). We found that participants with little evidence for base-rate neglect (i.e., S_{BR} values close to 1) were the ones for which the Bayesian model had been selected as the preferred model. Moreover, this analysis showed that the Skeptic Bayesian model and the Linear-Additive model competed for the remaining participants, with the relative success rate of each model being largely independent of the base-rate sensitivity value.

We next investigated the base-rate sensitivity values produced by the models themselves, based on the maximum-likelihood parameter estimates obtained above. The results (Figure 11B) show that both the Bayesian model and the Heuristic Toolbox model have difficulties in explaining low base-rate sensitivity values. In contrast, the remaining three models all seem able to produce a large range of base-rate sensitivities, including the Lexicographic model. Hence, the extremely poor performance of the Lexicographic model (see previous section) is not due to an inability to predict certain levels of base-rate neglect but must stem from other problems.

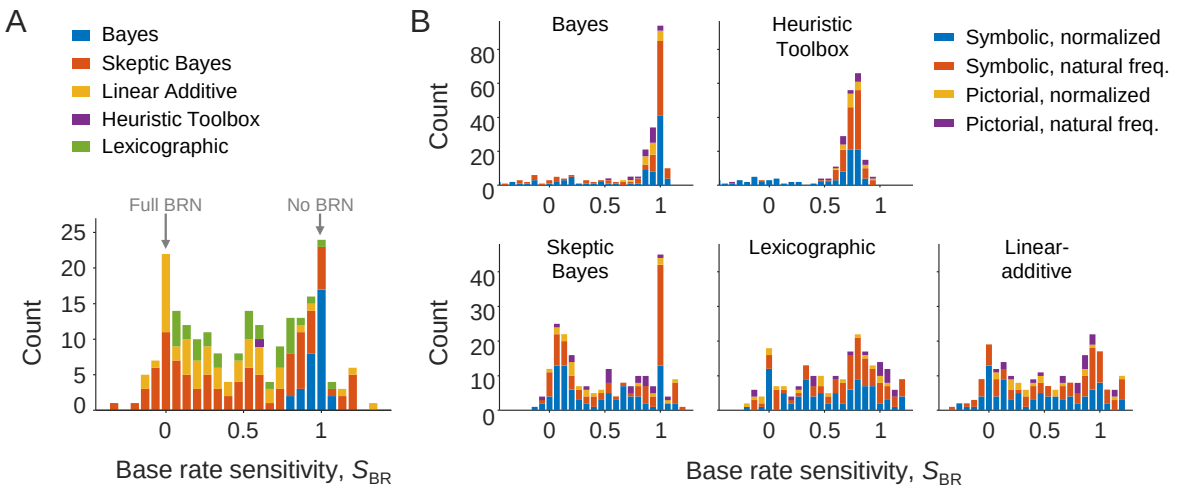

Figure 11 | Relation between decision strategies and base rate sensitivity. (A) The distribution of base rate sensitivity values (cf. Figure 4F), color-coded by the best-fitting model for each participant. Participants who showed little evidence of base-rate neglect (i.e., participants with S_{BR} values close to 1) are best accounted for by the Bayesian model. For the remaining participants, there is a competition between mainly the Skeptic Bayesian and Linear-Additive models, both of them able to account for a large range of sensitive values. (B) Model-based distributions of base-rate sensitivity values, computed from simulated responses under maximum-likelihood parameter estimates. Note that the graphs in the top row have a differenton the *y*-axis than those in the bottom row.

Frequency of heuristic decisions

The Heuristic Toolbox and Lexicographic models tested above incorporate a variety of heuristic decision rules ("Report the HR", "Report 1-FAR", etc). We found that both models provide relatively poor descriptions of the data. One possible explanation is that participants did generally not use the tested heuristics in their decision making. However, another possibility is that participants used a selection rule (which heuristic to use on which trial) that differed from the selection rules in the Heuristic Toolbox and Lexicographic models. To further investigate this, we counted for each participant in the two conditions with symbolic stimuli how many of their responses coincided with a heuristic decision. Even though not used in any of the tested models, we added "Report FAR" as an additional heuristic in this analysis. The results (Figure 12) showed that in both conditions approximately half of the responses were consistent with the disjunction of the five heuristics, while the other half were inconsistent with all five heuristics. Hence, even if we had a model to magically predict what specific heuristic will be used on a given trial – which we do not have – a heuristic-based model can never explain more than half of the responses in these data, unless additional processes are added.

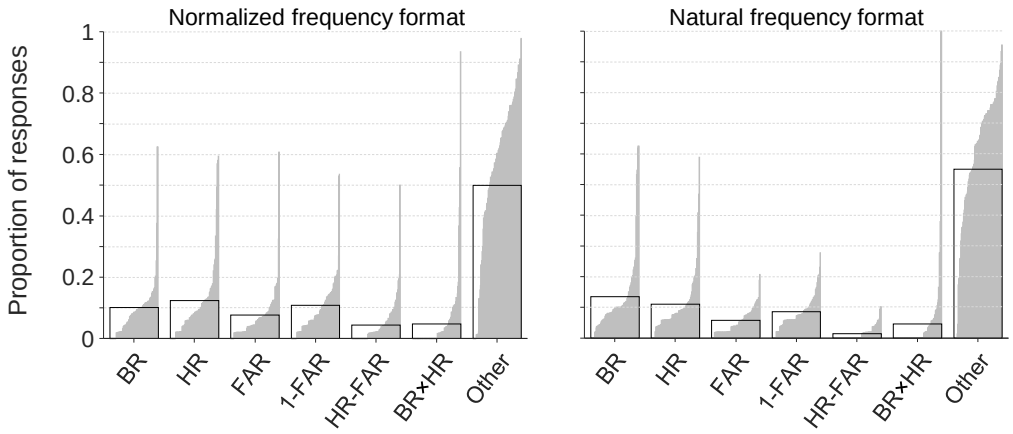

Figure 12 | Distribution of participant responses in the two conditions with symbolic stimuli. The proportion of participant responses that were equal (in two decimal places) to a heuristic compared to all other responses in the four conditions. Each bar corresponds to one participant. The black rectangles indicate the mean for each group of responses. In both conditions, approximately half of the responses coincided with one of the heuristics.

Carry-over effects

Since the participants performed multiple trials it is possible that they adjusted their responses based on their previous trials, which may have led to patterns in the data that might not had been there if they had performed only one trial. To examine whether this was the case we analyzed performance as a function of trial number in the MTurk data. These data had been collected in such a way that every item was presented as the first trial for at least two participants, as the second trial for two other participants etc. If performing multiple trials improved performance, the correlation between participants' responses and the correct responses should increase over trials. This was not the case (Figure 13). We also compared the *RMSD* of the first trial with that of all consecutive trials. For both the normalized- and natural frequency format, a one sample Bayesian *t*-test showed that the *RMSD* of the first trial is not larger than the rest (BF_{0-} = 19.70 and BF_{0-} = 61.86 in favor of the null hypothesis). Both these results suggest that the decision strategies employed by the participants were stable over trials, including the first one.

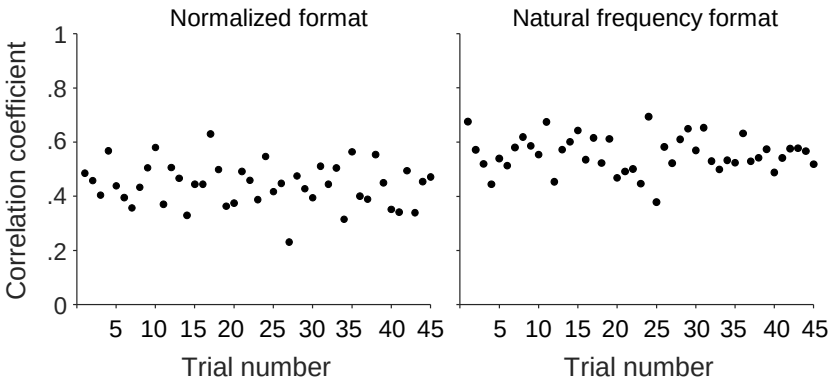

Figure 13 | Pearson product-moment correlation coefficients between the participants' responses and the correct responses, as a function of trial number. Correlations were computed separately for every trial number. Left: Normalized format. Right: Natural frequency format.

Discussion

In this study, we used a cognitive modelling approach to examine the strategies used by participants performing four versions of the classical medical diagnosis task. Unlike most studies using this task, we tested participants on a large range of different combinations of base rates, hit rates, and false alarm rates. Just as in previous studies, we found strong indications for base rate neglect, but with large individual differences in the severity of neglect and with a substantial number of participants not showing any signs of neglect. Moreover, we found indications that participants on average made proper use of the hit rate but underused the false-alarm rate (Figure 9). However, we also found strong individual differences in this respect. The individual differences were reflected in the model comparison results, which showed that there was not a single model that described the data well for all participants. While the Skeptic Bayesian was the preferred model for more participants than any other model, it was still the case that for about half of the participants one of the other models was preferred, in many cases by large margins. This heterogeneity warrants caution in drawing strong group-level conclusions about base rate neglect and indicates that this phenomenon is best studied at the level of individuals. Moreover, the success of the Skeptic Bayesian suggests that neglecting base rates might not be the main cause of people's reasoning errors; for a large number of participants, an alternative explanation could be that they reason in a way that resembles the application of Bayes rule, but affected by prior beliefs about base rates, hit rates, and/or false alarm rates. Since the incorporation of prior beliefs lies at the heart of Bayesian reasoning one could argue, somewhat ironically, that what has long been considered *the* Bayesian strategy to

solving the medical diagnosis task may be a mischaracterization of Bayesianism: it represents an observer who either has no prior experiences or completely ignores them.

Finally, even though the symbolic and pictorial tasks at first glance might seem rather different, the participants on average seemed to use the same kind of strategies to solve the tasks. This is an interesting finding since even though the underlying task structure was identical in the symbolic and the pictorial versions of the task, the judgments were based on exactly stated numbers in the symbolic condition and on uncertain assessments in the pictorial condition demanding the participants to reason under uncertainty.

Generality of the base-rate neglect effect

One of the main aims of this study was to examine whether previous reports of the base rate neglect phenomenon are specific to reasoning under extremely low base rates and high hit rates, or a more general property of human probabilistic reasoning. We found that signs of base rate neglect were present throughout the tested space of problems and to a similar degree as in the control experiment. Both in the control and in the main experiment there were two clusters of participants: those who seemed to largely ignore the base rate and those who seemed to account for it well. Particularly in the main experiment there was also a third group of participants who were spread relatively evenly between the two ends of the spectrum indicating large individual differences. In Figure 11 we can see that the Imperfect Bayesian accounts for the majority of participants clustered around $S_{BR} = 1$ while the Skeptic Bayesian and the Linear additive model account for the participants clustered close to $S_{BR} = 0$. This means that most of the participants who managed to make appropriate use of the base rates also integrated the information in accordance with Baye's theorem while the participants who neglected the base rate to a larger extent were best described with Skeptic Bayesian or Linear additive integration.

While the results from the control experiment replicated the findings from previous studies one must keep in mind that there were large individual differences regarding to what degree, if at all, an individual neglected the base rate. Since it is often only the averaged responses in the form of percentage correct or the modal response that is presented in papers it begs the question of how much of this individual variation is present also in previous studies who made strong claims of base-rate neglect. Due to the individual differences both in the control task and the main task of our study one needs to be careful about making population-level statements of base-rate neglect.

Cognitive mechanisms behind base-rate neglect: the skeptic Bayesian

Despite large individual difference, the skeptic Bayesian seemed to have a clear advantage in model selection over the other models. But what does it mean for a participant to be a skeptic Bayesian? For the participants who were insensitive to base rates (low S_{BR}-values in Figure 4F) it would mean that they had strong priors for the base rate and therefore did not base their estimate on the base rate given in the task. The estimated parameter values are however slightly puzzling (Figure 10): on average they are around 0.5 but the individual values are spread relatively evenly between 0 and 1. In terms of skeptic Bayesians, this means that the participants had very different prior beliefs regarding the prior of the base rate, hit rate and false-alarm rate. While this is entirely possible, it raises a red flag: if our participants had really been skeptic Bayesians with sensible prior beliefs, wouldn't we expect more homogeneous estimates in a pool of participants that presumably share very similar experiences?

If one were to take the evidence for the Skeptical Bayesian at face value, the interpretation would be that while people appear to combine the available information in a Bayesian way, they have wildly different beliefs regarding the prevalence of viruses and the sensitivity and specificity of medical tests. While their inference strategy is Bayesian one would be hard pressed to say that their decisions are in any way optimal.

Even though the skeptic Bayesian performs well in model comparison and could be a promising "middle ground" between perfect Bayesianism and heuristic-based decision making, it remains to be seen how viable of an explanation it is for base rate neglect. Since we used cross-validation, we deem the risk of overfitting as small, but there is still a possibility that there is another, yet-to-be-modeled mechanism underlying these participant decisions. While the Skeptic Bayesian may capture the main mathematical features of that mechanism, it could conceptually be very different.

Cognitive mechanisms behind base-rate neglect: linear additive integration

The model assuming linear additive integration accounts for the performance of a substantial portion of the participants. On this view, people typically have a qualitative understanding that both the evidence and base-rate is relevant in base-rate problems, but – as often observed in other multiple-cue judgments tasks – they spontaneously add up the cues, rather than engaging in the multiplication prescribed by probability theory (Juslin et al., 2009). This suggests that from long experience with the world, people may have normative insights at a general qualitative level, but they are unable to perform the normative mathematical integration when

presented with symbolic representations of probability. This interpretation is vindicated by the observation that, also after explicit tutoring and instruction on how to compute the posterior probability from Bayes' theorem in base-rate problems, people are better described by linear additive integration models than by the normative integration model (Juslin et al., 2011).

One potential criticism of the linear additive model concerns its flexibility; that its good fit is consistent with a wide array of apparently different strategies, ranging from exclusive weight on the hit-rate or base-rate, over the integration of pairs of components, to the weighting of all three components optimal for approximating Bayes' theorem. One could, however, argue that this accurately captures the richness and variety of the strategies used. Some participants indeed report the hit-rate directly (which is manifested by all weight on the hit-rate in the linear additive model); others take the hit-rate into account, but adjust it in view of the low base-rate (yielding non-zero weights for these components), whereas some participants appreciate the normative significance of all three components and integrate them as best they can. This may be determined by a variety of properties of the individuals, like their knowledge of probability, and triggered to a smaller and larger extent by various contextual cues as to what is relevant (Ajzen, 1977; Bar-Hillel, 1980; Birnbaum & Mellers, 1983; Fischhoff et al., 1979; Fishbein, 2015; Goodie & Fantino, 1999). The strategies are effectively identified by the parameters of the model and the common cognitive claim by the model is that people have difficulty with "number-crunching" symbolic representations of probability according to Bayes' theorem, but rather often default to linear additive integration of the components.

Heuristic models

The possible set of models one can choose to use in a study is in theory infinite. This is especially true for the heuristic models. The individual heuristic strategies (e.g., "multiply hit rate with base rate") used for modelling the data are not assumed to be used for all tasks but rather to be picked when the structure of the task at hand fits some criteria (e.g., "use joint occurrence heuristic if the base rate is high" (Gigerenzer & Hoffrage, 1995). However, the exact conditions for when the different heuristics should be used are difficult to establish and have therefore remained rather vague in the literature. For example, how high does the base rate has to be for it to be considered high? In this study we used two different heuristic models that circumvented this problem in different ways. The first model was based on a heuristic toolbox approach where we assume that the participant will use the most suitable heuristic on every trial. By looking at the best possible performance by this model (Figure 3) it became clear that

this heuristic model actually came very close to approximating the answer given by Bayes' rule, meaning that it is possible to use heuristics on this task and perform very well. The model comparison results provided little evidence, however, that this is a strategy used by the participants. One consequence of this approach is that participants who use heuristics in a way that is not well approximated by Bayes' rule will not be well fitted by the heuristic toolbox model. The second heuristic model was instead based on a lexicographic approach where people pick heuristics based on the perceived informativeness of the information in the task. From the model comparison results it was clear that this approach was also not used by many participants. While there may remain untested heuristic models with selection rules that differ from the ones we tested here, we could see from the distribution of responses in Figure 12 that it is still less than 50% of the answers that coincide with one of the heuristics. Part of these participants will also have used the strategy to consistently report only the base- or hit rate on every trial which means that they are best described by the linear additive integration model.

Effect of frequency format and presentation format

Using a natural-frequency format led to slightly better performance (a mean absolute error of 18 vs 23) and to more participants being classified as Imperfect Bayesians (21 vs 11). These changes were not as drastic as in many previous studies. The original study by Gigerenzer and Hoffrage found that changing from a normalized frequency format to a natural frequency format increased the performance rate from 16% correct responses to 46% correct responses (Gigerenzer & Hoffrage, 1995). A later study by Cosmides and Tooby found similar results, with a performance increase from 12% to 56% correct responses (Cosmides & Tooby, 1996). One explanation for the increase in the present study was smaller is that our participants were required to give their answer in percentages rather than a ratio. There is an important conceptual difference between asking for a ratio and asking for a probability in percentage, in the sense that the stochasticity is lost in the former case. If a frequency format helps people make estimations about stochastic events then the response should be in the form of a stochastic event and not in the form of a frequency. If a frequency format only helps people make estimations if the response is also given as a frequency then the conclusion should be that people are unable to make probability estimations. Another difference between our study and Gigerenzer and Hoffrage's is that our participants performed multiple trials with varying levels of BR, HR and FAR. However, to our knowledge there is no reason why this should have affected the beneficial effect of a natural frequency format.

A different explanation for why the natural frequency format was not as beneficial as in other studies is that the task structure did not make the set relations between BR, HR and FAR transparent enough (Barbey & Sloman, 2007; Sloman et al., 2003). The nested-sets hypothesis stem from the dual-process model and in contrast to Gigerenzer and colleagues' view that people are helped by the natural-frequency format in and of itself the proponents of nested-sets hypothesis argue that the format is only beneficial because it can make the task structure more transparent. In this view people use two systems to reason, one primitive associative judgment system that sometimes leads to errors in judgment and a second more deliberate rule-based system. The use of the second system is only induced if the task is represented in a way that is compatible with the rules. In this case the rules are elementary set operations which means that the problem needs to be readily formulated in terms of sets. The problem becomes even easier if the relevant sets are all nested (Sloman et al., 2003). In other words, if the chance of having the disease is nested within the chance of testing positive which in turn is nested within the set of all possible cases. Since in our task the HR≠1 it means that the sets are not nested (the chance of having the disease is not a subset of having a positive test) and therefore the task structure is less beneficial for the participants.

Even though the pictorial format introduced uncertainty in the task by forcing the participants to make their own estimations, this did not affect the overall performance. Looking at Figure 4B there is a trend towards better performance in the pictorial format but all in all there was anecdotal evidence against an effect of presentation format. This also means that the pictorial format did not function as a visual aid benefiting the participants as some previous studies has found (unless these two effects happened to cancel each other out). However, previous studies on the effects of visual representations have received mixed results and in the cases where a pictorial representation has been beneficial it was only presented as an addition to the question already using symbolic formulations (Brase, 2009; Garcia-Retamero & Hoffrage, 2013) and never by itself as was the case in our task. Why the two formats did not lead to any large differences in performance we can only speculate about; one intriguing possibility is that the reason the uncertainty in the pictorial task did not decrease the performance is that the mental representations of the symbolic numbers are just as uncertain. It is also possible that there are multiple factors working against each other. The task given in the symbolic format is similar to a typical math word problem and could therefore cause people with high levels of math anxiety to perform worse than they would have if they had received the pictorial task instead (Luttenberger et al., 2018).

Limitations and future directions

False alarm rate neglect. The fallacy of false-alarm-rate neglect has received support in previous studies. The current design was however mainly aimed at studying effect of base rates and not very suitable for studying sensitivity to hit rate and false alarm rate. The ranges of the hit rate and false alarm rate were shorter and with fewer levels than it was for the base rate. The reason for this was partly to decrease the number of trials needed for all stimulus configurations but mainly because the logic of the task should make sense. With the range of values that were used here the differences in the correct response was smaller and it was therefore difficult to evaluate whether participants made the correct adjustments or not based on sensitivity values as was done for the base rate in Figure 4F. What we can say is that based on the linear coefficients estimated for the participants classified as using linear additive integration, they did on average neglect the false-alarm rate (coefficient being close to zero) and made appropriate use of the hit rate (coefficient close to the optimal value, 0.52).

Using online surveys. Since there are a number of factors that differ between data collected in a lab versus when it is collected online via a service such as MTurk one has to be careful when comparing the results. Even though it is impossible to make sure that the data has been collected under similar conditions it is at least possible to compare the data after it is collected. In a separate data collection not reported on in this study the same exact task as the MTurk participants performed was performed by participants in the lab. These two data sets can therefore be used to investigate whether there are any systematic differences. In terms of performance, a Bayesian independent t-test showed anecdotal evidence in favor of there being no difference in the mean absolute error between the two groups (BF_{01}= 1.45). Another possible problem with using an online survey is the less controlled testing environment for the participants. Some participants performed very well and while they were explicitly told not to use any aids such as calculators, there is a possibility that they did. However, even if this was the case, they would still have to have known that Bayes' rule was the correct strategy even though they did not rely on their mental arithmetic ability to compute it.

Modeling. Although the set of models that we tested is larger than in most previous studies on base rate neglect, it was far from exhaustive and potentially interesting models remain untested. The large individual variation in explained variance suggest that for some participants we did indeed find a very likely candidate model but for others none of our proposed models provided a good fit. As discussed above, an important question raised by our modeling results is how we should understand the success of the skeptic Bayesian model: Do many people indeed reason as a Bayesian with strong prior beliefs about base rates, hit rate, and false-alarm rates?

Or does the skeptic Bayesian perhaps capture aspects in the data that happen to be mathematically similar to Bayesian reasoning with strong (and wildly varying) priors, but in reality, are grounded in a cognitive mechanism that is conceptually very different? Moreover, even though our results speak against the heuristic model that we tested here, it is of course possible that there are other heuristic models that fare better.

Within-subject testing of the benefits of the natural frequency format. Just as in previous work, we found a benefit of natural frequency format on performance. This effect has been explained by either referencing how during the evolution of the human race, naturally sampled frequencies is what we have been exposed to and subsequently evolved to use, or one has attributed the beneficial properties to the clarification of the task and thus making it easier for people to use deliberate thought instead of associative decision making. These two accounts are focused on how the difference in format makes people switch from towards using a different strategy. There is however also a possibility that the switch in format does not cause people to switch strategies but to use the same strategy with better tuned parameters. For example, a participant might shift from using a heuristic to a Bayesian strategy when the format changes from normalized frequencies to natural frequencies, or they might use a linear additive strategy in both formats but use better tuned weights. Our use of a between-subject design is a limitation in this respect. It would be interesting for future work to study the effect of frequency format using a cognitive-modelling approach applied to within-subject data. That would allow getting a more detailed insight into the nature of the reasoning strategies people use.

References

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. *Advances in Neural Information Processing Systems 30*, 1836–1846. https://doi.org/https://doi.org/10.1101/150052

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology*, *35*(5), 303–314. https://doi.org/10.1037/0022-3514.35.5.303

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

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*(3), 241–254. https://doi.org/10.1017/S0140525X07001653

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron*, *74*(1), 30–39. https://doi.org/10.1016/j.neuron.2012.03.016

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology*, *45*(4), 792–804. https://doi.org/10.1037/0022-3514.45.4.792

Bordalo, P., Gennaioli, N., & Shleifer, A. (2020). Memory, Attention, and Choice. *Quarterly Journal of Economics*, *135*(3), 1399–1442. https://doi.org/10.1093/qje/qjaa007

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, *23*(3), 369–381. https://doi.org/10.1002/acp.1460

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27. https://doi.org/10.1016/0030-5073(74)90002-6

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica*, *87*(2–3), 137–154. https://doi.org/10.1016/0001-6918(94)90048-5

Brookes, B. C., Simon, P., de Laplace, M., Truscott, F. W., & Emory, F. L. (1953). A

Philosophical Essay on Probabilities. *The Mathematical Gazette*, *37*(320), 148. https://doi.org/10.2307/3608971

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by Physicians of Clinical Laboratory Results. *New England Journal of Medicine*, *299*(18), 999–1001. https://doi.org/10.1056/nejm197811022991808

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

Costello, F., & Watts, P. (2014). Surprisingly rational: probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480. https://doi.org/10.1037/a0037010

Costello, F., & Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, *89*, 106–133. https://doi.org/10.1016/j.cogpsych.2016.06.006

Costello, F., & Watts, P. (2018). Probability Theory Plus Noise: Descriptive Estimation and Inferential Judgment. *Topics in Cognitive Science*, *10*(1), 192–208. https://doi.org/10.1111/tops.12319

Crupi, V., Tentori, K., Lombardi, L., Cherubini, A. M., Evans, J., Fitelson, B., Girotto, V., Legrenzi, P., & Osherson, D. (2009). Pseudodiagnosticity Revisited. *Association*, *116*(4), 971–985. https://doi.org/10.1037/a0017050

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*. https://doi.org/10.1016/0010-0277(92)90049-N

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*(1), 180–209. https://doi.org/10.1037/0033-295X.106.1.180

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., & Koechlin, E. (2016). Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, *0*(0), e1003661-10313. https://doi.org/10.1016/j.neuron.2016.11.005

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processes. *Psychological Review*, *101*(3), 519–527. https://doi.org/10.1037/0033-295X.101.3.519

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews. Neuroscience*, *9*, 292–303. https://doi.org/10.1038/nrn2258

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, *50*(2), 123–129. https://doi.org/10.1007/BF00309212

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance*, *23*(3), 339–359. https://doi.org/10.1016/0030-5073(79)90002-3

Fishbein, M. (2015). Progress in social psychology. In *Progress in Social Psychology* (Vol. 9). Psychology Press. https://doi.org/10.4324/9781315725765

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science and Medicine*, *83*, 27–33. https://doi.org/10.1016/j.socscimed.2013.01.034

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Gigerenzer, G., & Todd, P. M. (1999). Fast and Frugal Heuristics: The Adaptive Toolbox. In *Simple Heuristics That Make Us Smart* (pp. 3–34). https://doi.org/10.1177/1354067X0171006

Goodie, A. S., & Fantino, E. (1999). What does and does not alleviate base-rate neglect under direct experience. *Journal of Behavioral Decision Making*, *12*(4), 307–335. https://doi.org/10.1002/(SICI)1099-0771(199912)12:4<307::AID-BDM324>3.0.CO;2-H

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x

Harris, A. J. L., Corner, A., & Hahn, U. (2009). Estimating the probability of negative events. *Cognition*, *110*(1), 51–64. https://doi.org/10.1016/j.cognition.2008.10.006

Harris, A. J. L., De Molière, L., Soh, M., & Hahn, U. (2017). Unrealistic comparative

optimism: An unsuccessful search for evidence of a genuinely motivational bias. *PLoS ONE*, *12*(3). https://doi.org/10.1371/journal.pone.0173136

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin*, *138*(2), 211–237. https://doi.org/10.1037/a0025940

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. https://doi.org/10.1016/0010-0285(92)90002-J

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*(1), 259–298. https://doi.org/10.1016/j.cognition.2007.02.003

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*, 856–874. https://doi.org/10.1037/a0016979

Juslin, P., Nilsson, H., Winman, A., & Lindskog, M. (2011). *Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats*. https://doi.org/10.1016/j.cognition.2011.05.004

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, *10*(3), 189–209. https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<189::AID-BDM258>3.0.CO;2-4

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): a “lazy” algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*(5), 563–607. https://doi.org/10.1207/s15516709cog2605_2

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*.

Khan, A., Breslav, S., Glueck, M., & Hornbæk, K. (2015). Benefits of visualization in the Mammography Problem. *International Journal of Human Computer Studies*, *83*, 94–113. https://doi.org/10.1016/j.ijhcs.2015.07.001

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, *13*(6), 493–497. https://doi.org/10.1016/S0960-9822(03)00135-0

Lopes, L. L. (1985). Averaging rules and adjustment processes in Bayesian inference. *Bulletin of the Psychonomic Society*, *23*(6), 509–512. https://doi.org/10.3758/BF03329868

Luttenberger, S., Wimmer, S., & Paechter, M. (2018). Spotlight on math anxiety. In *Psychology Research and Behavior Management* (Vol. 11, pp. 311–322). Dove Medical Press Ltd. https://doi.org/10.2147/PRBM.S141421

McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, *143*(12), 1273–1312. https://doi.org/10.1037/bul0000126

McKenzie, C. R. M. (1994). The Accuracy of Intuitive Judgment Strategies: Covariation Assessment and Bayesian Inference. *Cognitive Psychology*. https://doi.org/10.1006/cogp.1994.1007

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*(3), 194–216. https://doi.org/10.1037/h0048070

Nilsson, H., Juslin, P., & Olsson, H. (2008). Exemplars in the mist: The cognitive substrate of the representativeness heuristic: Cognition and Neurosciences. *Scandinavian Journal of Psychology*, *49*(3), 201–212. https://doi.org/10.1111/j.1467-9450.2008.00646.x

Oaksford, M., & Chater, N. (1994). A Rational Analysis of the Selection Task as Optimal Data Selection. *Psychological Review*, *101*(4), 608–631. https://doi.org/10.1037/0033-295X.101.4.608

Ofir, C. (1988). Pseudodiagnosticity in judgment under uncertainty. *Organizational Behavior and Human Decision Processes*, *42*(3), 343–363. https://doi.org/10.1016/0749-5978(88)90005-2

Sanborn, A. N., & Chater, N. (2016). Bayesian Brains without Probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. https://doi.org/10.1016/j.tics.2016.10.003

Shanteau, J. (1975). Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica*, *39*(1), 83–89. https://doi.org/10.1016/0001-6918(75)90023-2

Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology*. https://doi.org/10.1016/0010-0285(75)90006-7

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological*

Bulletin, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, *91*(2), 296–309. https://doi.org/10.1016/S0749-5978(03)00021-9

Stengard, E., & Van Den Berg, R. (2019). Imperfect Bayesian inference in visual perception. *PLoS Computational Biology*, *15*(4). https://doi.org/10.1371/journal.pcbi.1006465

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. In *Science* (Vol. 331, Issue 6022, pp. 1279–1285). https://doi.org/10.1126/science.1192788

Tversky, A., & Kahneman, D. (1981). *Evidential Impact of Base Rates*.

van Opheusden, B., Acerbi, L., & Ma, W. J. (2020). Unbiased and efficient log-likelihood estimation with inverse binomial sampling. *PLOS Computational Biology*, *16*(12), e1008483. https://doi.org/10.1371/journal.pcbi.1008483

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*(JAN), 1. https://doi.org/10.3389/fnins.2012.00001

Zhu, J. Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian Sampler: Generic Bayesian Inference Causes Incoherence in Human Probability Judgments. *Psychological Review*. https://doi.org/10.1037/rev0000190