2023-06

Heterogeneity of Rules in Bayesian Reasoning: A Toolbox Analysis

Woike, JK

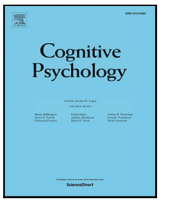
10.1016/j.cogpsych.2023.101564
Cognitive Psychology
Elsevier

Cognitive Psychology 143 (2023) 101564

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych



Heterogeneity of rules in Bayesian reasoning: A toolbox analysis

Jan K. Woike [a,b,*], Ralph Hertwig [a], Gerd Gigerenzer [a]

[a] Max Planck Institute for Human Development, Center for Adaptive Rationality (ARC), Lentzeallee 94, 14195 Berlin, Germany
[b] University of Plymouth, School of Psychology, Portland Square, Plymouth PL4 8AA, UK

ARTICLE INFO

Keywords:
Bayesian reasoning
Cognitive modeling
Model competition
Process heterogeneity
Simple rules
Interindividual differences

ABSTRACT

How do people infer the Bayesian posterior probability from stated base rate, hit rate, and false alarm rate? This question is not only of theoretical relevance but also of practical relevance in medical and legal settings. We test two competing theoretical views: single-process theories versus toolbox theories. Single-process theories assume that a single process explains people's inferences and have indeed been observed to fit people's inferences well. Examples are Bayes's rule, the representativeness heuristic, and a weighing-and-adding model. Their assumed process homogeneity implies unimodal response distributions. Toolbox theories, in contrast, assume process heterogeneity, implying multimodal response distributions. After analyzing response distributions in studies with laypeople and professionals, we find little support for the single-process theories tested. Using simulations, we find that a single process, the weighing-and-adding model, nevertheless can best fit the aggregate data and, surprisingly, also achieve the best out-of-sample prediction even though it fails to predict any single respondent's inferences. To identify the potential toolbox of rules, we test how well candidate rules predict a set of over 10,000 inferences (culled from the literature) from 4,188 participants and 106 different Bayesian tasks. A toolbox of five non-Bayesian rules plus Bayes's rule captures 64% of inferences. Finally, we validate the Five-Plus toolbox in three experiments that measure response times, self-reports, and strategy use. The most important conclusion from these analyses is that the fitting of single-process theories to aggregate data risks misidentifying the cognitive process. Antidotes to that risk are careful analyses of process and rule heterogeneity across people.

In Laplace's famous conjecture, "probability theory is nothing but common sense reduced to a calculus" (1814/1951, p. 1969). One might assume that once psychology was established as an academic field, psychologists would have begun to examine actual probabilistic reasoning and test Laplace's conjecture. But that did not happen. Before 1950, probability theory played virtually no role in the psychological research on reasoning, neither as a model of how people ought to reason nor of how people do reason (Gigerenzer & Murray, 2015). Not until the middle of the 20th century did this change, with Brunswik's (1955) probabilistic functionalism and (Piaget and Inhelder's, 1951/1975) treatise *The origin of the idea of chance in children*. Neither, however, made any reference to Bayes's rule. Psychologists only began to focus on Bayes's rule as a model of human inference with the advent of Edwards's research program (e.g., Edwards, 1968).

Largely consistent with Laplace's conjecture, (Edwards, 1968) concluded that "opinion change is very orderly and usually proportional to numbers calculated from Bayes's theorem" (p. 17). In his view, people were conservative Bayesians, that is, they reason in a Bayesian way but accord too much weight to base rates. This view was overturned in the 1970s by Kahneman and Tversky, who concluded that "man is apparently not a conservative Bayesian: he is not Bayesian at all" (1972, p. 450). They proposed that people instead employ the representativeness heuristic, thereby judging probability by similarity. Since then, other descriptive alternatives to Bayes's rule have been proposed. Juslin et al. (2009) argued that people weigh and add probabilities

* Corresponding author at: University of Plymouth, School of Psychology, Portland Square, Plymouth PL4 8AA, UK.
E-mail address: jan.woike@plymouth.ac.uk (J.K. Woike).

https://doi.org/10.1016/j.cogpsych.2023.101564
Received 21 May 2021; Received in revised form 2 March 2023; Accepted 31 March 2023
Available online 11 May 2023

instead of multiplying them, as assumed in Bayes's rule. Although conservatism, the representativeness heuristic, and the weighing-and-adding model portray the cognitive processes in very different ways, they share one assumption, namely that people's reasoning in Bayesian inference tasks is best modeled in terms of a single process plus error. Single-process theories assume, at least implicitly, homogeneity in the cognitive process, meaning that people diverge from Bayes's rule in a single systematic way. For conservatism the deviation is small and in the direction of base rate information; for the representativeness heuristic, the deviation is substantial and in the opposite direction of conservatism; for the weighing-and-adding model, in contrast, the deviation can be small or large in either direction, depending on what parameters (e.g., weights) are chosen.

According to single-process theories, human reasoning in Bayesian inference tasks essentially fails in a uniform manner. This indeed echoes findings in numerous reasoning tasks beyond Bayesian inference. For instance, in the bat-and-ball problem, most people commonly get it wrong and always in the same way until they are exposed to the correct answer (Frederick, 2005; Woike, 2019). In the Wason selection task, which tests the validity of if-then rules, most select the p and q cards instead of the p and not-q cards (Wason, 1966). And in the Linda task, most people deem it more probable that Linda is a bank teller and active in the feminist movement than that she is a bank teller (Tversky & Kahneman, 1983). Reasoning errors are thus often seen as analogous to visual illusions that trigger the same erroneous perceptions across individuals. For instance, in Shepard's (1990) table illusion, a majority of people see two tables as systematically different although their length and width are identical. In visual and cognitive illusions, so the common view, distortions are systematic and consistent across people. In contrast to this view, we argue that in Bayesian inference tasks people err in at least five distinct ways. This is consistent with Brunswik's (1955) argument that errors are large and diverse in cognitive tasks, and small and unimodal in perceptual tasks.

In this article we proceeded in five steps. First, we ask whether past studies indeed support single-process theories of Bayesian reasoning by testing whether the distribution of individual inferences is unimodal, as predicted by a single process. Using a convenience sample of studies, we demonstrate that the distributions are anything but unimodal. Second, we ask how then have influential conclusions about single-process theories in the literature emerged? One potential reason is that inferences produced by multiple rules can nevertheless be best fitted by a single-process theory. To analyze this possibility, we simulated individuals using multiple rules responding to Bayesian reasoning tasks. We show that a single-process theory, specifically, the weighing-and-adding model, nevertheless achieves the best fit, and surprisingly, the best out-of-sample prediction of the aggregate. Consequently, cross-validation does not prevent misidentifying the process. Third, we turned to the question of what toolbox of rules people draw upon in Bayesian inference tasks. To this end, we analyzed the actual individual responses of 4,188 participants to 106 Bayesian inference tasks. Five non-Bayesian rules plus Bayes's rule (henceforth referred to as the *Five-plus toolbox*) can precisely predict the responses of about two thirds of over 10,000 inferences. This conclusion is based on an analysis of the posterior probability judgment only. Fourth, in three experiments we provide independent evidence for the Five-plus toolbox by making use of response times as well as self-reported information and strategy use. Finally, we show that when the Five-plus toolbox is used to predict a distribution of judgments, it outperforms single-process theories, which mischaracterize the process.

1. Single-process theories of Bayesian reasoning

According to the single-process view, the errors people make in Bayesian inference tasks are unimodal. We consider three single-process theories and define them as follows:

1.1. Conservatism

Inspired by Edwards (1968) and adapted to the task format considered here, conservatism can be formalized as

$$\hat{p}(H|D) = w \cdot p(H) + (1 - w) \cdot \frac{p(H)p(D|H)}{p(H)p(D|H) + p(\neg H)p(D|\neg H)} + \epsilon, \tag{1}$$

with w (ranging between 0 and 1) representing a weighting parameter that determines the relative impact of the base rate on the estimate $\hat{p}$ and ϵ representing an unsystematic error. If $w = 0$, Eq. (1) is identical to Bayes's rule.

1.2. Representativeness heuristic

As originally proposed by Kahneman and Tversky (1972), we use a common quantitative definition of "representativeness" from the literature (Gigerenzer & Murray, 2015):

$$\hat{p}(H|D) = p(D|H) + \epsilon, \tag{2}$$

with ϵ representing an unsystematic error.

1.3. Weighing-and-adding model

We took the weighing-and-adding model from Juslin et al. (2009). It is formally defined as

$$\hat{p}(H|D) = \alpha + w_b \cdot p(H) + w_h \cdot p(D|H) + w_f \cdot p(D|\neg H) + \epsilon, \tag{3}$$

with w_b, w_h, w_f representing parameters that determine the relative impact of base rate, hit rate, and false alarm rate, and α an additive constant. Estimates for the parameter values were taken from Juslin et al.'s "behavioral additive model" (p. 869). The parameter ϵ represents an unsystematic error.

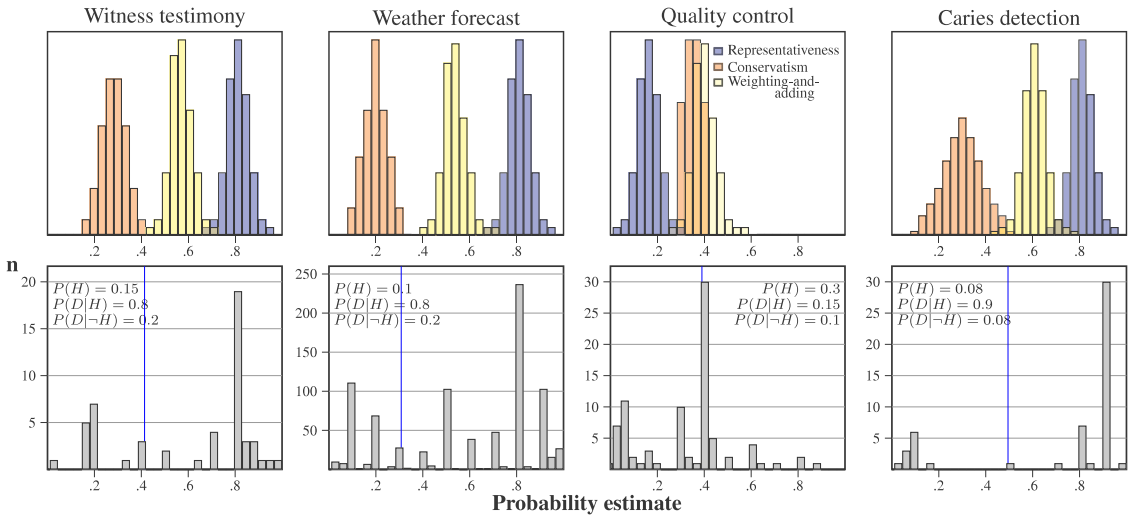


Fig. 1. Predicted and empirical distributions of inference of Bayesian posterior probabilities in four different studies (left to right): Bar-Hillel (1980; Cab task, $N = 52$ prospective students); Dohmen et al. (2009; Weather forecast, $N = 859$ representatively sampled German residents); Hoffrage et al. (2015; Quality control, $N = 110$ managers and students of management), and Nadanovsky et al. (2018; Caries detection, $N = 52$ dentists). The upper panels show the pattern of predicted inferences for each of the single-process theories: conservatism, representativeness heuristic, and weighing-and-adding (Eqs. (1)–(3)). The lower panels show the empirically obtained distributions of estimates (one per participant).

2. Study 1: Do the distributions of inferences conform to single-process theories?

One prediction that follows from the single-process theories are distinct unimodal distributions (due to unsystematic error) of individual inferences. To test this prediction, we took advantage of four studies reporting individual data. They were conducted across four decades, from 1980 to 2018, and represent a range of participant populations (i.e., samples of students, citizens, MBA students and managers, and dentists) and countries (Israel, Germany, Switzerland, and Brazil). The tasks used in these studies are the classic witness testimony task, also known as the cab task (Bar-Hillel, 1980), and three more recent Bayesian inference tasks: the weather forecast task (Dohmen et al., 2009), the Skiwell quality control task (Hoffrage et al., 2015), and the caries detection task (Nadanovsky et al., 2018). Details on these tasks can be found in Appendix A.

The predictions of the three single-process theories are shown in the top panel in Fig. 1. Each predicts a unimodal distribution, but in different regions of the x-axis. To derive the predictions, we used Equations (1) to (3). We used illustrative values for ϵ for the representativeness heuristic and the weighing-and-adding model. For the conservatism model, we set $w = 0.5$, and chose ϵ so that predictions covered the interval between base rate and the Bayesian response. The predictions thus illustrate the expected qualitative distribution pattern (see Study 6 for an in-depth exploration of these parameters). The four bottom panels show the actual observed distributions of individual estimates. Predicted and empirical distributions differ systematically. Specifically, all empirical distributions have multiple peaks rather than one, they show discontinuity rather than an approximately continuous distribution, and they range across the entire probability scale rather than being centered on a narrow range. Consider, for illustration, the cab task, which has several peaks: the highest at 80%, and the second highest, far removed, at 20%, with a total range from 5% to 98%. The same discrepancy between predicted and observed error distributions also occurs in the three other Bayesian tasks.

With respect to actual inference in the tasks, these four studies consistently reveal heterogeneity, discontinuity, and a large range in values. Each of the three single-process theories, in contrast, implies a unimodal distribution and fails to explain the observed range and discontinuity in the inferences. These characteristics of the distributions are not limited to the four studies in Fig. 1. We observed the same heterogeneity in the 106 Bayesian tasks analyzed below (see Fig. 5). This triad of heterogeneity, discontinuity, and range raises the question of why single-process accounts are (still) the dominant account in the literature. Specifically, in light of these empirical distributions, how is it possible that single-process theories have been shown to fit the data best (e.g., Juslin et al., 2009)? We offer a possible explanation by demonstrating that parameterized single-process theories can fit and predict inferences best. In the next section, we show that this holds even if the inferences are generated by a group of individuals who draw upon a heterogeneous set of rules, with each individual consistently using one of the rules, and if none of the rules corresponds to the single process.

3. Study 2: The amazing flexibility of the weighing-and-adding process

To analyze the fitting and predictive performance of single-process models, we simulated a response distribution of 200 individuals relying on a heterogeneous set of rules when responding to 40 Bayesian inference tasks. Having created the simulated respondents and their inference rules, we know exactly which rule generated each inference. Each respondent consistently used

Table 1
Definition of Bayes' rule and three non-Bayesian rules used in the simulation, and the number of cues used by each rule. b = base rate or prior probability; h = hit rate; f = false alarm rate.

	Rule	Formula	Cues used
(1)	Bayes's rule	$P(H\|D) = \frac{b \cdot h}{b \cdot h + (1-b) \cdot f}$	3
(2)	Joint occurrence	$P(H\&D) = b \cdot h$	2
(3)	Likelihood subtraction	$P(D\|H) - P(D\|\neg H) = h - f$	2
(4)	Base-rate only	$P(H) = b$	1

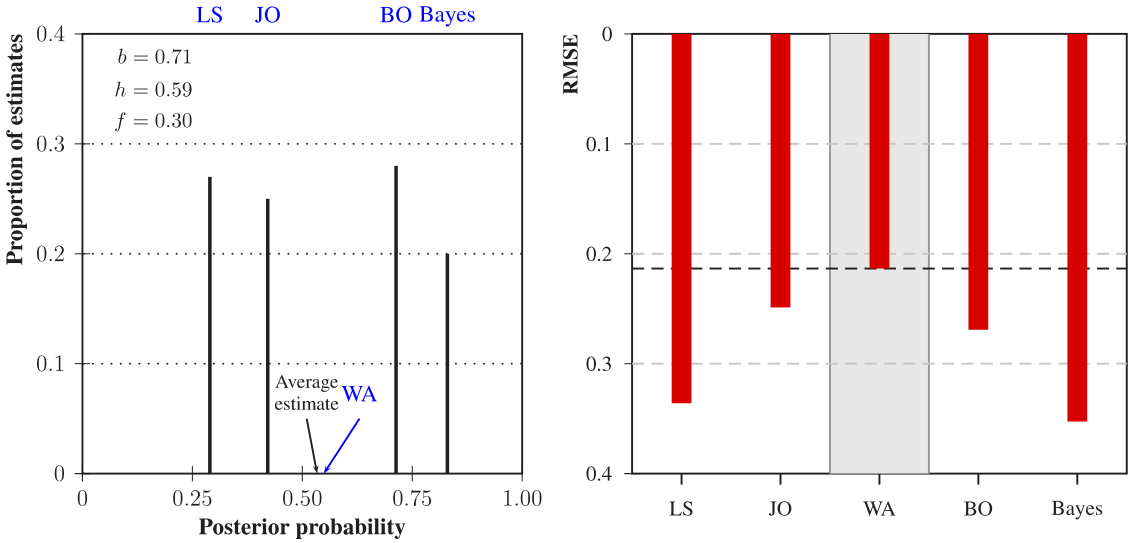


Fig. 2. A parameterized single-process model (weighing-and-adding, Eq. (3)) has the lowest root mean squared error (RMSE) in predicting the inferences of simulated individuals, none of whom used weighing-and-adding. Left panel: The posterior probabilities provided by 200 simulated respondents in one of the 40 generated Bayesian tasks. The proportion of estimates for each of the four generating rules is shown, as is their average. The parameters of the weighing-and-adding model (WA) were estimated based on half of the dataset; its predictions in the other half of the data set are plotted. Right panel: RMSE for the five strategies (four rules plus WA) for the same task. LS = likelihood subtraction; Bayes = Bayes's rule; BO = base-rate only; JO = joint occurrence.

the same rule for each task, but rules differed across people. To arrive at a realistic composition of the repertoire of rules, we implemented Bayes's rule, joint occurrence, likelihood subtraction, and the base-rate only rule (defined in Table 1). These rules are among the most frequently used non-Bayesian rules in Gigerenzer and Hoffrage's (1995) analysis and include rules considered and studied by Juslin et al. (2009, p. 869).

We constructed 40 Bayesian tasks in which the base, hit, and false alarm rates were randomly drawn from a uniform distribution, with the realistic constraints that the hit rates were greater than 50% and the false alarm rates less than 50%. Each simulated respondent was randomly assigned to one of the four rules with equal probability, so that each rule was used approximately equally often across the population. Fig. 2 illustrates for one of the 40 tasks the resulting heterogeneous distribution of inferences.

We first fitted the parameters of the weighing-and-adding (WA) model (Eq. (3)) to half of the data generated across 40 Bayesian tasks. Fig. 2 (left panel) shows its prediction in the other half for the illustrative task along with the average of all participants' inferences; it also plots the proportion of inferences by the generating rules. Importantly, although the WA model does not predict a single individual inference (generated by the four rules), it closely approximates the average of the generated inferences. Next, we calculated the root mean squared error (RMSE) for each of the four generating rules and the WA model. The RMSE for the latter is the smallest, meaning that it outperforms each of the generating rules (see Fig. 2, right panel). This analysis of one illustrative Bayesian task replicates the previously reported finding that the WA model provides the best fit to the average data. Interpreting this fit as evidence that the model is the most appropriate theory of the processes (e.g., Juslin et al., 2009, 2011) is, however, highly questionable. Our example also shows that this fit can be achieved irrespective of the fact that the model fails to capture a single individual inference yet well accommodates the average inference produced by a toolbox of varied rules.

A basic insight from research on model selection is that models with free parameters need to be tested in out-of-sample predictions, not just by data fitting alone. The WA model has four free parameters (see Eq. (3)), whereas each of the four rules in Table 1 has none. Thus, the WA model is more flexible in fitting. Yet the result in Fig. 2 is based on out-of-sample prediction, not fitting. The result may instead be a chance result due to the structure of the illustrative Bayesian task we chose. Therefore, we repeated the demonstration in Fig. 2 with the total set of 40 tasks and tested the WA model in out-of-sample prediction, asking: Will the single-process model still come out ahead? Specifically, we created 1,000 populations of 200 simulated respondents each; each respondent received all 40 tasks created for each population. We used the same four generating rules as in Fig. 2 and the same random allocation of respondents to rules. Next, we fitted the WA model on half of the inferences in each population and then tested

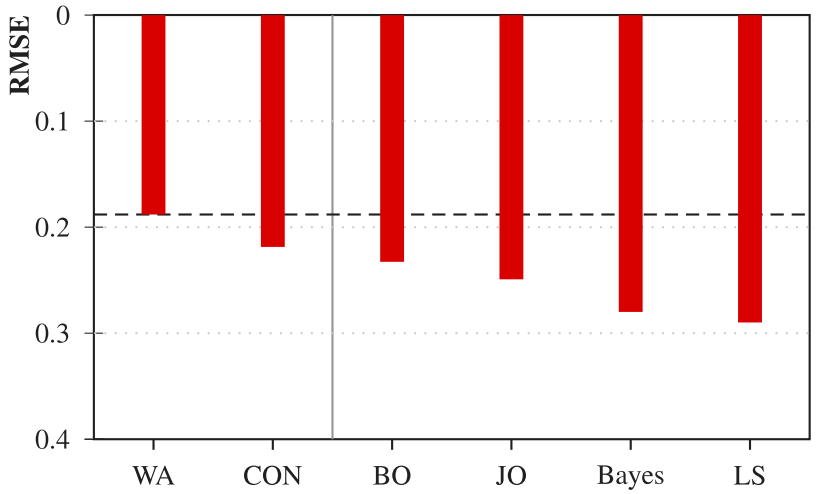


Fig. 3. The out-of-sample predictive power of the weighing-and-adding (WA) model that is not employed by any of the simulated respondents. Shown is the average RMSE for each rule across 1,000 populations of 200 simulated respondents for all 40 tasks. Parameters for the WA model and the conservatism (CON) model were estimated based on half of the responses in each population. LS = likelihood subtraction; Bayes = Bayes's rule; BO = base-rate only; JO = joint occurrence.

it on the other half. To compare its performance with another single-process model, we also implemented a conservatism model by estimating its parameter on the same half of inferences (using constrained regression). We repeated this 100 times for each of the 1,000 populations. As Fig. 3 shows, the WA model's RMSE was again better than that of any of the generating rules. Furthermore, it was able to produce the best prediction of the average inferences, even though it again could not predict a single respondent's inferences better than chance.[1] The conservatism model came close to the performance of the WA model, with the second-lowest error of all models. The important lesson from this analysis is that even if a single-process model such as the WA model is tested in out-of-sample prediction (cross-validation), this still does not protect researchers from drawing erroneous conclusions when the generating process represents a toolbox of varied rules.

When the focus is turned to modeling an individual's inferences, the seductive power of the WA model does not wane either, as the next analysis demonstrates. We took half of the inferences of each of the simulated individuals, fitted the parameters of the WA model and the conservatism model, and then predicted the other half of the inferences separately for each individual. We observed not only an even better RMSE for the WA model but also an excellent score of correct predictions. Fig. 4 plots these findings. This stunning performance once again results from the model's high flexibility. It is so underspecified that two of the four generating rules are special cases of the model: base-rate only ($w_b = 1$, $w_h = 0$, $w_f = 0$, $\alpha = 0$, $\epsilon = 0$), where all weights except that for the hit rate are zero, and likelihood subtraction ($w_b = 0$, $w_h = 1$, $w_f = -1$, $\alpha = 0$, $\epsilon = 0$). As Fig. 4 illustrates, these two rules make distinct predictions and implement different cognitive processes. The conservatism rule is also able to match the predictions of Bayes's rule ($w = 0$) and base-rate only ($w = 1$), and it shows a similar performance in terms of percentage correct. Its RMSE is again held back by its limited flexibility.[2] To conclude, even in situations in which one knows that the distribution of judgment was produced by four specific rules, a single-process model with free parameters can outperform those generating rules in fitting the average judgments, in out-of sample predictions of average judgments, and in out-of-sample predictions of individual judgments.

4. Study 3: Which rules are in the toolbox?

The results of Study 1 suggest that several rules are inter- and possibly intra-individually employed in Bayesian inference tasks. Which rules are these? In order to answer this question, we collected 30 existing and original data sets comprising 106 tasks and 4,188 human participants, amounting to 10,562 inferences (see Appendix B). To the best of our knowledge, this is the largest set of inferences in Bayesian tasks ever analyzed. Our first goal was to identify the candidate rules that people employ. To this end, we examined the literature for studies identifying non-Bayesian rules and found a total of 44 candidate rules (Cohen & Staub, 2015; Gigerenzer & Hoffrage, 1995; Macchi, 2000; McKenzie, 1994; Mellers & McGraw, 1999; Zhu & Gigerenzer, 2006, and Studies 4a and 4b in this article). The complete set of rules and their policies are described in the Supplementary Material (Table S10).

Once the candidate rules were identified, we used them to make predictions for each rule (including Bayes's rule) across the 106 tasks. We then tested how well these rules and their predictions fared across the 30 empirical data sets. To avoid misclassifications,

[1] As a control, we also tested the WA model in a setting in which responses were generated by two further rules, representativeness and the false alarm complement rule (Juslin et al., 2009). The model again had the smallest RMSE in out-of-sample prediction.

[2] We demonstrate this relative limitation further in another set of simulations (see Supplementary Material, Section 1.4) after replacing base-rate only by the representativeness rule.

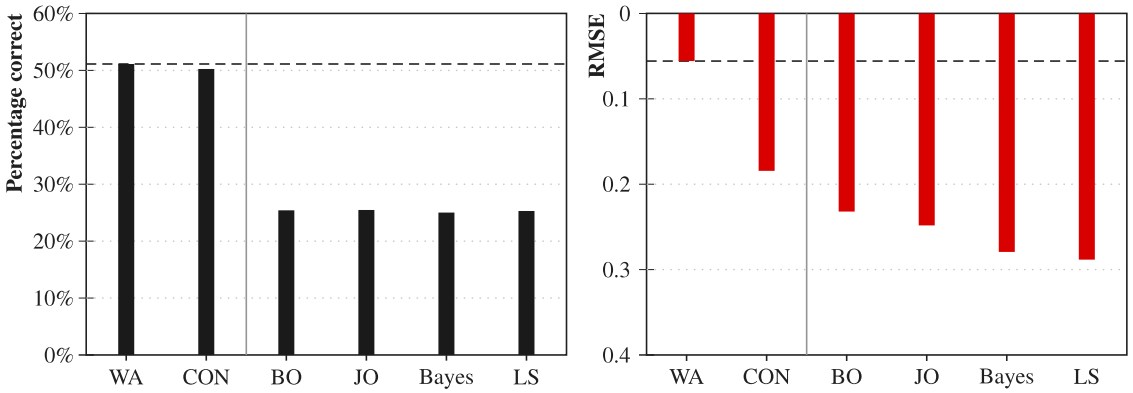


Fig. 4. Illustration of the individual-level out-of-sample predictive power of a parameterized WA model. Results for the five rules across 1,000 populations of 200 simulated respondents with 40 tasks; individual parameters for the WA model and CON model were estimated based on half of the responses for each individual. The left panel shows the average proportion of correct predictions, the right panel the RMSE for each of the rules. The dashed lines mark the best observed performance for each criterion (WA in both cases).

Table 2
The most common rules found in the literature across 106 Bayesian inference tasks, ordered by how well they predict individual responses (median across tasks: Median). Also shown is the percentage of responses across all individuals and tasks (Average), and the potential of the free parameters of the weighing-and-adding rule to mimic a rule (Mimic). The abbreviations for the rules are explained in Table 1, except for REP = representativeness, FC = false alarm complement, and 50% = 50% rule.
Source: C: Cohen and Staub (2015); G: Gigerenzer and Hoffrage (1995); J: Juslin et al. (2009); K: McKenzie (1994); Ma: Macchi (2000); Me: Mellers and McGraw (1999); Z: Zhu and Gigerenzer (2006); S4: Studies 4a and 4b in this article.

Rank	Rule	Abbreviation	Source	Median	Average	Mimic
1	h	REP	C, G, K, Me, S4, Z	24.0	23.8	Yes
2	b	BO	C, G, Ma, Me, S4, Z	11.6	20.5	Yes
3	$1-f$	FC	C, G, Ma, S4	11.1	14.7	Yes
4	$h-f$	LS	G, Ma, S4	8.8	11.1	Yes
5	$b \cdot h$	JO	G, Me, S4	6.7	8.8	No
6	0.5	50%	Me, S4	3.1	9.0	Yes
7	$b \cdot (h+f)$	–	S4	2.5	8.2	No
8	$\frac{b \cdot h}{b \cdot h + (1-b) \cdot f}$	Bayes	All	2.4	3.9	No
9	f	–	C, Me, S4	2.4	5.8	Yes
10	$\frac{h+f}{2}$	–	S4	1.1	6.3	Yes
11	$0.5 \cdot b$	–	S4	1.1	5.6	Yes
12	$\frac{b \cdot h}{(1-b) \cdot f}$	–	S4	1.1	2.4	No
13	$\frac{h}{h+f}$	–	K,Ma	1.0	8.9	No
14	$b \cdot h + (1-b) \cdot f$	–	Me, Z	0.8	2.7	No
15	$h-b$	–	G	0.6	3.2	Yes

we classified an inference as consistent with a rule only if empirical and predicted response was numerically identical or within ± 0.1%. Exceptions to this strict criterion were permitted solely for rules in which the process of multiplication was invoked; here we granted a rounding to the closest percentage (some interfaces that were used in experiments did not allow for a higher degree of precision). For some of the 106 tasks the predictions of different rules could overlap. In these cases, more than one rule was considered to be correct, which led to an inflated performance estimate for some of the matching rules (e.g., with a base rate of 0.35, a hit rate of 0.75, and a false alarm rate of 0.4, both the BO rule (b) and the LS rule ($h-f$) determine 0.35 to be the answer; a participant's judgment of 0.35 would thus be counted as a correct prediction for both rules). To counteract these chance hits, we ordered the rules by their median performance per task, given that the rules' estimates did not overlap in more than a few of the environments (see also the list of tasks in the Supplementary Material, Table S9). Rules that predicted 0% of cases (median) were excluded from further analysis. Table 2 lists the 15 analyzed rules and Fig. 5 plots the median proportion of correct predictions across tasks and the proportion across all judgments for these rules and the weighted adding models.

The predictions of Bayes's rule match few estimates, with a median performance across tasks of only about 2.4% of responses (average = 3.9%). Among the non-Bayesian rules, the representativeness heuristic scored best (Eq. (1)), with a median of 24.0% of all inferences (average = 23.8%), followed by the base-rate rule with 11.6% (average = 20.5%). In third place is the false alarm complement (1 minus false-alarm rate; equivalent to the specificity of the test), which scored a median of 11.1% correct predictions (average = 14.7). In fourth place is likelihood subtraction (hit rate minus false alarm rate), scoring 8.8% correct predictions (average = 11.1%). Coming in fifth, is the product of base rate and hit rate, also known as joint occurrence, with about 6.7% correct predictions (average = 8.9%). Each of these five non-Bayesian rules accounted for at least a median of 5% of estimates. The median

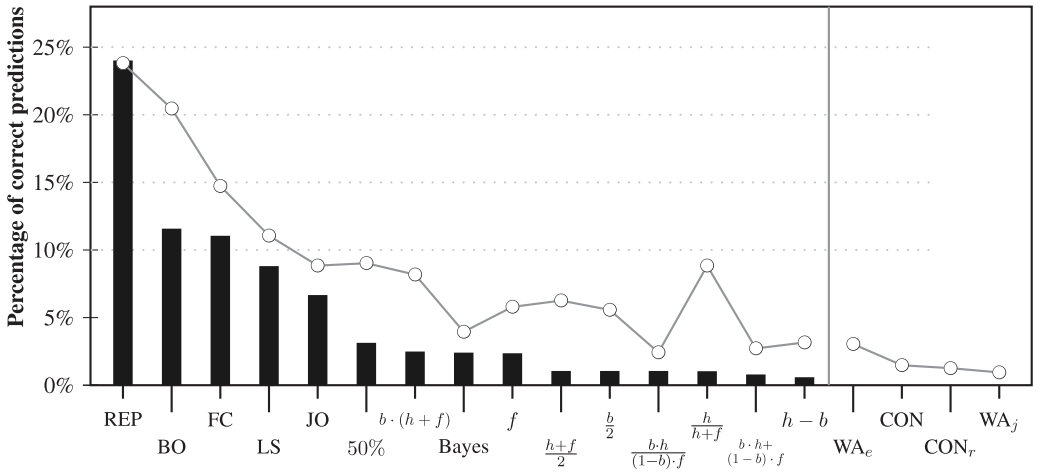


Fig. 5. Correct predictions of 15 rules in Table 2: Bars show the median percentage across 106 tasks; circles show the average percentage across all judgments. All models based their estimates on base rate (b), hit rate (h), and false alarm rate (f) as specified on the x-axis (the 50% rule always predicted a posterior probability of 0.5, regardless of values). Bayes = Bayes's rule. There are two weighing-and-adding rules estimated based on two different empirical data sets. The parameter values for WA_e were estimated from all 10,562 responses in the 106 tasks ($a = 0.148$, $w_b = 0.316$, $w_h = 0.435$, $w_f = -0.128$), while WA_j used parameter values from Juslin et al. (2009, p. 869; $a = .18$, $w_b = .61$, $w_h = .46$, and $w_f = -.4$). The parameters for CON ($w = 0.092$) and CON_r ($w = 0.680$) were estimated using the same procedure. The median percentage of correct predictions was 0% for all of these four models.

match of all other rules is markedly lower. The set of the five rules plus Bayes's rule (the *Five-plus toolbox*) account for a total of 64.3% of all inferences.[3]

We also considered the performance of the three single-process theories proposed (see Eqs. (1)–(3)). As mentioned, there is clear evidence for the representativeness heuristic—but in terms of one of the rules in the toolbox rather than in terms of a single-process theory. The WA_e rule, with parameters fitted on the entire empirical dataset, predicts less than 4% of estimates. An even more dismal performance (less than 1%) holds for its variant, the WA_j, with parameters fitted in Juslin et al.'s studies (2009, 2011). Finally, we tested two versions of conservatism, fitting the parameter based on the full dataset (CON) or restricted to responses that fell in the interval between the Bayesian solution and the base rate (CON_r). The performance of these models fell between the performance of the two weighing-and-adding rules. The median match (across tasks) of all four single-process models was zero.

As a control, we also conducted this analysis on the individual level. Depending on the experiment, participants responded to between one and 36 tasks, with a median of two per participant. What proportion of each individual participant's inferences can be precisely predicted by at least one rule in the set of the five non-Bayesian rules and Bayes's rule? Recall that none of these six rules uses free parameters. For 2,292 (55%) of the 4,188 participants, each single inference they made could be explained by one of the five rules or by Bayes's rule (Fig. 6). For 2,917 (70%) participants, all or some of their inferences could be predicted by one of the five rules or by Bayes's rule. A further 168 participants (4%) answered with 0.5 in all tasks, corresponding to the 50% heuristic. The inferences made by a total of 239 participants can be better explained by including this strategy. The WA rule added little in predicting individual participants' inferences. For six participants (0.1%), the only rule predicting at least one of their inferences was WA. For 20 additional participants (0.5%), a few more judgments could be predicted with WA than without it (but other rules explained at least some of their inferences).

We present additional evidence for intra-individual consistency in the Supplementary Material. In separate simulations, we replicated the individual-level findings presented for artificial data in this article (see Fig. 4) using two empirical datasets that offered sufficient responses from each individual participant. We further demonstrate that a model that identified a participant's most likely rule (constrained to the parameter-free rules in the toolbox) had a good predictive performance and outperformed a parameterized weighing-and-adding model (see SM, section 1.3).

In conclusion, the analysis of this comprehensive dataset of people's Bayesian inferences suggests three key results. First, there is general evidence for heterogeneity in participants' inference rules. Second, this heterogeneity fails to support single-process theories even though there is evidence for the representativeness heuristic (Eq. (1)) as one of several rules rather than as a single-process theory. Third, a set of five non-Bayesian rules and Bayes's rule explains every inference of 55% of the respondents, and a total of 64% of all inferences across all respondents. This is corroborated by additional individual-level simulations (see SM, section 1.3). These results indicate that individuals tend to use one of the rules in the toolbox consistently rather than switching between rules.

[3] This percentage was calculated without relying on the table data. It is smaller than the sum of the percentages across the six rules, as—in several tasks—some inferences were accounted for by more than one rule. The percentage is close to the sum of median percentages that are better protected against such outliers.

Prediction	REP	BO	JO	LS	FC	Bayes
All estimates	18.8	17.6	5.5	3.9	10.4	2.1
Some estimates	13.0	13.0	10.2	10.7	8.7	4.0

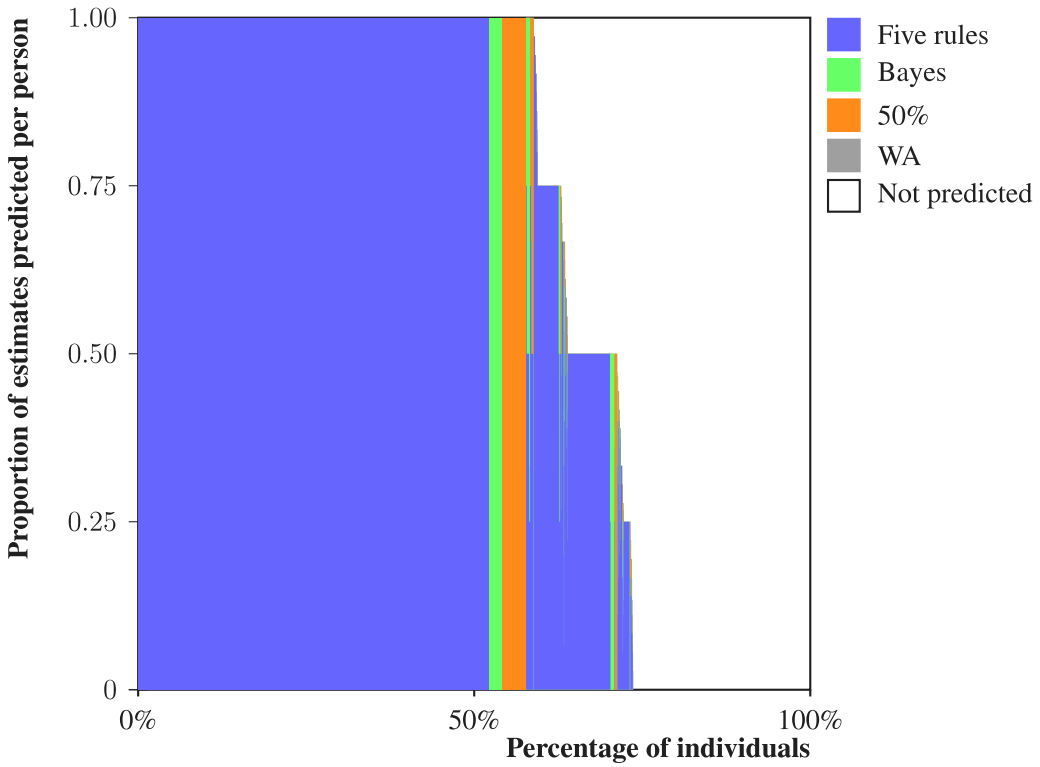


Fig. 6. Individuals consistently tend to use a single rule from the toolbox. Shown are the inferences predicted by the five rules (REP, BO, JO, LS, FC), Bayes's rule, the 50% rule, and the weighted additive rule (from left to right). Each of the 4,188 individuals corresponds to a vertical slice, ordered by the intra-individual consistency in rule use. On the left side are individuals who relied on any one of the five rules in every task, followed by those who relied on Bayes's rule and on the 50%-rule in every task. Further to the right are individuals who relied on the five rules in less than 100% of tasks – and did not rely on Bayes's rule or the 50%-rule exclusively – in order of decreasing match with all considered rules. For 55% of all participants, every judgment can be explained by the set of the five rules and Bayes's rule. The table above the diagram shows the percentage of participants for whom the six individual rules predict all of their estimates and the percentage for whom each rule predicts some but not all of their estimates.

5. Study 4: Is there process evidence for non-Bayesian rule use?

In order to further test the validity of the Five-plus toolbox, we conducted three experiments measuring response times and self-reported solution strategies.[4] Studies 4a–c measured response times. In Studies 4a and 4b, participants were asked which values (cues) they used to arrive at an estimate in the Bayesian inference task; in Study 4c, they had to click on the information they had used.

5.1. Study 4a

5.1.1. Sample

Participants were recruited on Amazon Mechanical Turk. The Bayesian inference tasks were a part of a larger survey – offered as a "Human Intelligence Task" (HIT) – estimated to take between 15 and 20 min for most participants and offered a fixed payment of USD 1.25, with a performance-contingent bonus of up to USD 1.00. Participants were required to be located in the US and to have a minimum approval rating of 95% and a minimum number of completed HITs of 50. If multiple surveys started from the

[4] The actual probability judgments, but not the reaction times and the self-reports, had been included in Study 3. See the SM, sections 2–4 for the study materials, section 5 for additional results.

same IP address, only the first attempt was analyzed if it did not overlap with the second. This resulted in a final sample size of 545 participants (250 male, 295 female, Mean age= 36.1 years).

5.1.2. Method

Participants responded to four Bayesian inference tasks and then to a questionnaire. Each participant saw the same four tasks and the same four number sets (specifying base rate, hit rate, and false alarm rate), with a systematically varied assignment of number sets to texts and a systematically varied order of tasks between participants. The number sets correspond to tasks 402, 501, 502, and 504 in Table S9 (see SM). Response times were recorded for each task. After the final task, participants were given a general account of Bayesian inference tasks that explained which elements were named base rate, hit rate, and false alarm rate. On the basis of this explanation, participants were asked to indicate which numbers they had used in calculating solutions, to pick out the rules they used out of a list of rules, and to answer questions about their attendance of statistics classes, their attitude towards calculation, and an incentivized estimate of their performance relative to others. Last but not least, participants were invited to describe employed rules not yet specified in the list in an open answer format.

5.2. Study 4b

5.2.1. Sample

In Study 4b, 319 participants were recruited on Amazon Mechanical Turk under the same conditions as in Study 4a (141 male, 178 female, Mean age= 36.5 years).

5.2.2. Method

Participants responded to two Bayesian inference tasks and to the same questionnaire as in Study 4a. Each participant saw the same two texts and two out of four number sets (specifying base rate, hit rate and false alarm rate), with a systematically varied assignment of number sets to texts and a systematically varied order of tasks between participants. The number sets correspond to tasks 601–604 in Table S9. Response times were recorded for each task.

5.3. Study 4c

5.3.1. Sample

In Study 4c, participants were recruited on Amazon Mechanical Turk. The Bayesian inference tasks were part of a larger HIT estimated to take between 10 and 15 min for most participants and offered a fixed payment of USD 1.50, with a performance-contingent bonus of up to USD 0.20. Participants were required to be located in the US and to have a minimum approval rating of 96%. In addition, participants had to correctly answer two out of three attention check questions to begin the task and pass checks preventing participation from non-US locations or via the use of virtual private servers or virtual private networks (Burleigh et al., 2018). If multiple surveys started from the same IP address, only the first attempt was analyzed if it did not overlap with the second, resulting in a final sample size of 1,013 participants (538 male, 474 female, Mean age= 36.7 years).

5.3.2. Method

Participants responded to one Bayesian inference task. Each participant saw the same text and one out of eight sets of numbers, specifying base rate, hit rate, and false alarm rate. The number sets correspond to tasks 901–908 in Table S9. Response times were recorded. We chose number sets so that the rules that had been observed most frequently resulted in unique values. After submitting an inference, participants were shown an image map containing the text and were asked to highlight the numbers they had used in calculating their solution by a mouse click. The numbers, corresponding to base rate, hit rate, false alarm rate, and two distractors (irrelevant for solving the task), were individually selectable.

5.4. Results of studies 4a–c

Participants were classified as using a specific rule if one and only one rule corresponded to at least half of their estimates (two in Study 4a, one in Studies 4b and 4c). Studies 4a and 4b probed a total of 864 participants to also report directly which of the three values in the task (base rate, hit rate, false alarm rate) they used to come up with an answer. Assuming that people have some insight into their rule use, one would expect some level of consistency between the value(s) reported and the rule that is consistent with the numerical judgment.

Table 3 (left) shows that among those classified as users of the base-rate rule, 92% reported having used the base rate. Among those classified as users of the representativeness heuristic (REP), 88% reported having used the hit rate. This consistency between rule classification and reported information use also held for the three more complex rules: Bayes theorem, joint occurrence, and likelihood subtraction. Finally, using the more intuitive method of directly indicating used cues via mouse click, the same results could be replicated in Study 4c (Table 3, right).

The time needed to respond varied systematically between rules, with those whose responses corresponded to Bayes's rule clearly taking the longest time (see Supplementary Material, Section 5.5).

Table 3
Percentage of participants who reported having considered a cue by observed rule use in Studies 4a and 4b (left side) and percentage of participants who clicked on a cue value in Study 4c (right side). The three cues were base rate (b), hit rate (h), and false alarm rate (f). As a control, Study 4c introduced two distractor numbers x_1 and x_2 in the text; these had no bearing on the correct solution. Bolded numbers correspond to the values necessary to compute the respective rule. The rightmost column in each block reports the number of participants classified.

Strategy		Studies 4a–b				Study 4c					
		% considered			N	% considered					N
		b	h	f		b	h	f	x_1	x_2	
Bayes	$\frac{b \cdot h}{b \cdot h + (1-b) \cdot f}$	**93**	**97**	**93**	30	**100**	**100**	**100**	0	0	7
JO	$b \cdot h$	**92**	**100**	25	24	**70**	**91**	22	0	0	23
LS	$h - f$	42	**83**	**71**	24	23	**85**	**73**	0	4	26
FC	$1 - f$	42	66	**76**	38	25	41	**84**	3	3	32
REP	h	56	**88**	31	186	5	**90**	7	1	1	206
BO	b	**92**	38	21	191	**87**	5	5	2	2	391
Total		74	72	44	861	51	44	28	3	4	1,010

6. Study 5: The seductive power of the single-process illusion

In Figs. 2 and 3, we showed that highly parameterized models such as the WA model could create the illusion of a single process predicting people's inferences. This illusion results from the use of RMSE for model testing in situations of rule heterogeneity. RMSE is based on the assumption of unimodal distributions, as in a single-rule model. Additionally, we showed that out-of-sample testing is no remedy. In a next step, we demonstrated that the same illusion arises when the empirical data of the 4,188 human participants and 10,562 inferences are analyzed. Splitting the inferences in the empirical dataset randomly into two halves, we used one half to estimate the parameters for the weighing-and-adding model. We then measured the performance in the other half. This process was repeated 10,000 times. Once again, when RMSE is used, the WA model achieves the best value. However, when percentage correct is used as performance criterion, the RMSE winner cannot predict people's inferences beyond chance level (see Fig. 7). The discrepancy between the WA model's impressive performance in the RMSE analysis and its dismal performance in percentage correct can be best understood by its logic, demonstrated in Fig. 2, of betting on the middle of the road even if nobody takes this path. This procedure minimizes squared deviations from actual inferences but fails to predict the actual process.[5] As another demonstration of the concurrent success and vacuity of this logic, we replaced the four-parameter WA model with, perhaps, the simplest rule: Always respond with 50% (cf. Fig. 6). This rule also hits the middle of the response scale. A test on the entire data set with over 10,000 judgments showed that it achieved a RMSE practically identical to that of the weighing-and-adding model and even predicted 9% of participants' judgments, more than were predicted by the weighing-and-adding-model.

7. Modeling heterogeneity of rules from a toolbox point of view

Our analysis of an extensive data set found strong evidence that people use a toolbox of rules in Bayesian inference tasks. We also showed that no single-process model is able to explain people's inferences. Yet a single-process theory, the WA model, achieves excellent fitting and prediction performance, regardless of the fact that it entirely misconstrues the actual processes. How should one respond to this methodological challenge? We next spell out a procedure of steps in order to examine and model heterogeneity in rule use and to protect against the seductive power of single-process theories.

Step 1. Check for heterogeneity: Examine whether the response distribution for the same task is unimodal or multimodal. This means going beyond the analysis of aggregate values. A multimodal distribution is one where distinct values on the range of the dependent variable occur with relatively high frequency and are clearly separated from each other. An ideal case of heterogeneity would be observed if distribution were found to be both multimodal and non-continuous.

Step 2. Hypothesize cognitive processes that imply the modes of the distribution: On the basis of an analysis of the literature (and existing data sets), of an analysis of the cognitive competences that a population of individuals are likely to have (e.g., a person lacking any formal education in probability theory should not be expected to be able to compute the Bayesian answer in a task that represents information in terms of probabilities) or, alternatively, of process tracing methods (e.g., MouseLab, Johnson et al., 1989; think-aloud, Ericsson and Simon, 1984; EEG; eye tracking, Schulte-Mecklenbeck et al., 2011), develop hypotheses about the rules that produce the systematical part of the heterogeneity in the response distribution. Define these rules precisely, with a minimum of free parameters.

Step 3. Test the set of candidate processes: Rely on data not employed in Step 2 to test for the existence and prevalence of the candidate rules. Do not use RMSE for model testing because it assumes a single mode with a symmetric error distribution. Use percentage correct instead (see Fig. 5).

Step 4. Analyze the ecological rationality of the individual rules: This requires specifying the structure of environment in which one rule, relative to another, is reasonable in light of the ecological conditions. For instance, if the quality of the new data is questionable, relying on the base rate can be a reasonable strategy.

[5] The SM, section 1.1, defines and tests a model based on simple rules that mimics the performance of the WA model.

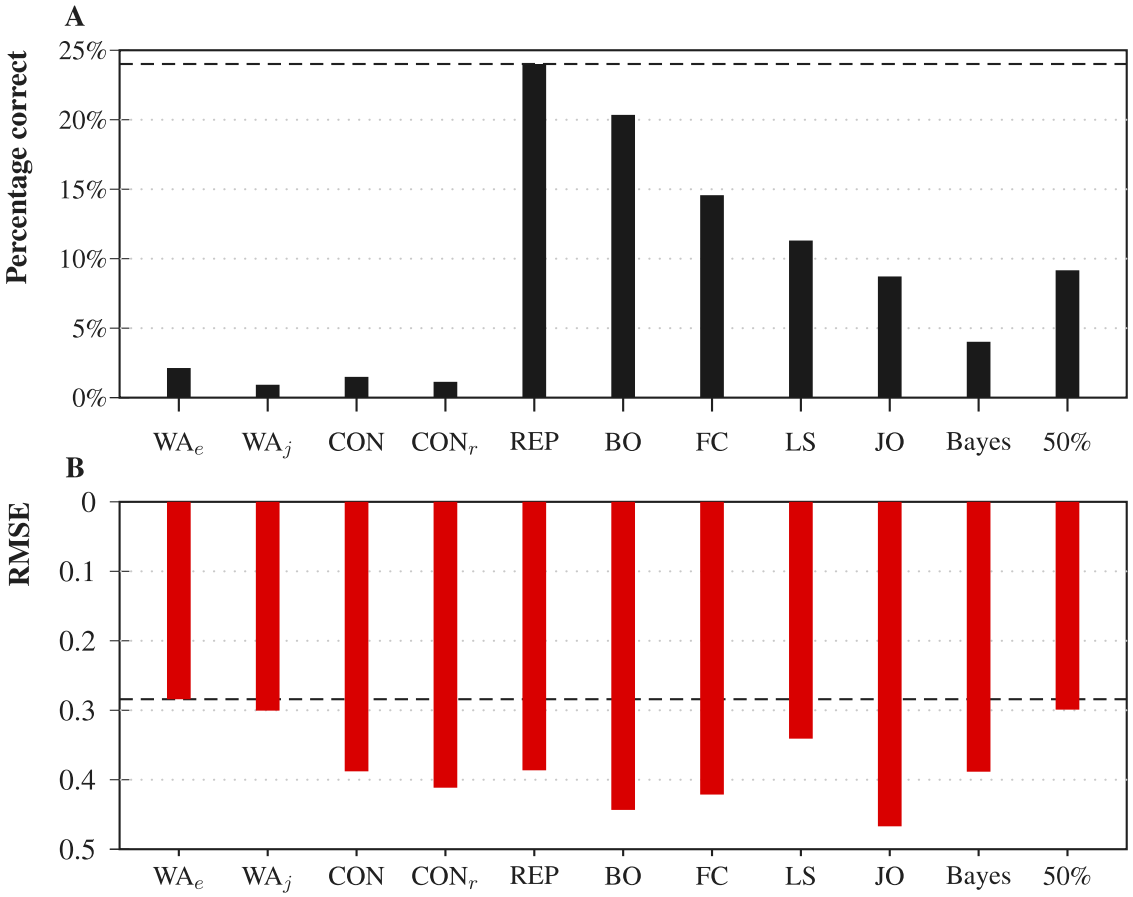


Fig. 7. The single-process model (WA) has the lowest error, as measured by RMSE, for the empirical dataset of 4,188 individuals, similar to the simulated data (Fig. 3). At the same time, its percentage of correct predictions is close to zero. Shown are the average results across 10,000 trials: (A) average percentage correct for each rule, (B) root mean squared error (RMSE) for each rule on an inverted y-axis. WA$_e$ corresponds to the empirical weighted additive model based on the data set, WA$_j$ to the weighted additive model with parameter values taken from Juslin et al. (2009). CON and CON$_r$ are the two conservatism models with parameter values based on the full dataset and filtered responses, respectively. The dashed lines correspond to the best observed performance for each criterion.

Step 5. Define a toolbox of rules and conditions of their successful use: On the basis of such a toolbox, it is possible to examine whether and to what extent people's actual use of the rules obeys conditions of ecological rationality (e.g., Goldstein & Gigerenzer, 2002; Horn et al., 2016; Woike et al., 2017).

In what follows, we have implemented Steps 1 to 3 (but not Steps 4 and 5).

7.1. The five-plus toolbox model

The logical answer in the case of heterogeneity is to abstain from making single-point predictions and instead predict distributions of responses. The "Five-Plus toolbox model" is a toolbox theory that postulates that the heterogeneity in Bayesian inferences can largely be captured by five rules plus Bayes's rule, each of which has a characteristic proportion p_i $(i = 1, \ldots, 6)$. Its predicted response distribution ($\hat{RD}$) can be captured as follows:

$$\begin{aligned}\hat{RD} = \{&prediction(REP), p_1; prediction(BO), p_2; prediction(FC), p_3;\\ &prediction(LS), p_4; prediction(JO), p_5; prediction(Bayes), p_6\}'\end{aligned} \tag{4}$$

with $\sum_{i=1}^{6} p_i = 1$.

The proportions of how frequently each rule is employed can be estimated out-of-sample by randomly dividing the individual responses into two sets, using the first for estimation and the second for testing the model. If p'_1 to p'_6 are the observed proportions, parameters are estimated as

$$p_i = \frac{p'_i}{\sum_{i=j}^{6} p'_j}. \tag{5}$$

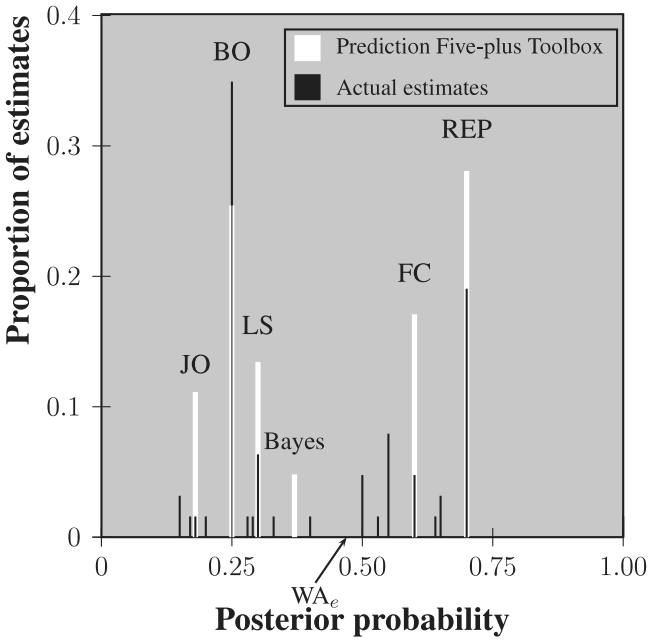


Fig. 8. Predictions of the Five-plus toolbox model for a single task. The predicted proportions are shown in white bars; the empirical data are shown in black bars. Also shown is the prediction of the empirical weighted additive model (WA_e).

This permits the predicted proportions for each rule to be compared with the actual proportions instead of relying on RMSE. Fig. 8 illustrates this procedure using the empirical data for one of the 106 tasks (see Study 3).

Fig. 8 demonstrates the predictive accuracy, defined as average overlap of predicted and observed proportions summed across 101 intervals,[6] that is,

$$Predictive\ accuracy = \sum_{i=1}^{101} min(observed_i, predicted_i). \tag{6}$$

Average Overlap is thus a specific measure of percentage correct. The out-of-sample predicted proportions of the Five-plus toolbox model are shown by six white bars. Fifty-seven percent of both distributions overlap, which means that the Five-plus model correctly predicted 57% of the actual distribution. The model underestimated the reliance on the base-rate only rule, and overestimated the reliance on the other five rules. The empirical weighing-and-adding model, in contrast, could not predict a single observed response.

Next, we repeated the procedure for the entire empirical dataset analyzed in Study 3. We estimated the parameters of the Five-plus model for half of the 10,562 individual responses and then tested the predicted proportions on the other half of the responses. This procedure was repeated across 10,000 different splits. The average predictive accuracy was 59.3%. As a control, we repeated that same procedure by splitting the 106 tasks into two halves, using one half to estimate the parameters and the other to measure the predictive power (averaged across 10,000 different splits). The result was practically identical (59.8%; see Fig. 9), suggesting that individuals' rules are fairly constant across tasks, a result consistent with Fig. 6. As another control, we used the same procedure to show the predictive accuracy of the empirical weighing-and-adding model.[7]

In conclusion, the Five-plus toolbox model correctly predicted the distribution of 59%–60% of the individual responses. Note that this percentage refers to all responses, not all of which are predictable, and is a strict test of the model. In comparison, the weighing-and-adding model achieved 2% only.

7.2. Weighing-and-adding with context-dependent weights?

One possible response defending the weighing-and-adding model against the presented evidence may focus on two assumptions that underlie our analysis: We assumed that weighing parameters are (1) the same across participants, and (2) the same across tasks. Dropping both assumptions simultaneously would render the model virtually useless, as any pattern of individual results could be explained by a combination of task-specific and individual-specific weights. In addition, our analysis of process data showed no evidence for such enormous model and weight flexibility. Our analysis of tasks with multiple repeated measures (see SM, section 1.3)

[6] Of these, 99 intervals were 1% wide, the two outer intervals 0.5% wide ([0, 0.5%[and [99.5; 100%]).

[7] The SM, section 1.2, reports results for toolbox models with different numbers of contained rules.

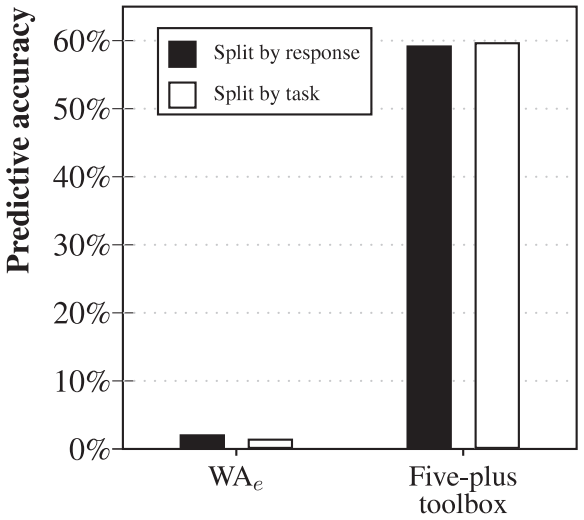


Fig. 9. Predictive accuracy of the Five-Plus toolbox for 10,562 individual responses to a total of 106 Bayesian tasks. The entire set of individual responses was split into two halves, the parameters of the Five-Plus toolbox model were estimated on one half, and the parameterized model was tested on the other half (black bars). As a control, the split was made on tasks, half of which were used for fitting the parameters, and the other half for testing. In each case, 10,000 splits were made. Shown are the average proportion of correct predictions, defined as the proportion of actual inferences predicted by the model.

demonstrated that the heuristics in the toolbox outperformed an *individualized* weighing-and-adding model. In fact, weighing-and-adding matched the performance of the toolbox model only for those participants whose responses were consistent with heuristics that represent special instances of the weighing-and-adding equation. There was no evidence for individualized weights. We address the promise of task-specific weights in Study 6 when asking the broader question whether there is even a possibility – given our data – for a single-process model to outperform the proposed toolbox, even when we allow for error in individual predictions and task-specific parameters.

8. Study 6: Is there any single-process model that can outperform the toolbox?

The present studies highlight an important methodological point. The evaluation of a model by RMSE can systematically deviate from its evaluation by percentage correct if there is heterogeneity in the cognitive processes. We argued that under heterogeneity RMSE is likely a potentially misleading criterion of success because it wrongly assumes a unimodal distribution of responses. Still, one could object that one should always use the same criterion for model development and model evaluation, regardless of heterogeneity or homogeneity in process.

Specifically, using percentage correct as model evaluation criterion might put single-process models at a disadvantage as long as they are bound to use RMSE for parameter estimation, as it was done for weighing-and-adding in our studies.[8] To rely on RMSE for parameter estimation is the very standard in linear models such as weighing-and-adding. Nevertheless, let us analyze a scenario in which one estimates the parameters by maximizing percentage correct. In this case, one could expect weighing-and-adding to equal, or at least, to approximate the performance of the most frequently employed heuristic it is able to mimic. This, however, implies that there is no longer any weighing and adding in the eponymous model. It is reduced to the very heuristic it mimics. Study 6, however, yields an even stronger and more general result. No single-process model – irrespective of whether its parameters are derived on the basis of RMSE, percentage correct or any other quality criterion – can reach the performance of the toolbox model. This finding is not restricted to fitting and out-of-sample predictions. It even holds if one handicaps the toolbox model by allowing the single-process models to take on task-dependent parameter values while the toolbox model operates with fixed parameters (see also SM, section 6.6). These analyses show that, in the presence of heterogeneity, the toolbox model outperforms single-process models regardless of whether model development and model testing are based on matching criteria or not.

In order to be able to arrive at this strong result, we first established upper limits for the performance of any conceivable single-process model. Specifically, we gave every advantage to the single-process model: access to the data set used to test the performance, as well as the opportunity to optimize parameters in order to maximize the target overlap criterion. Furthermore, these models were tested and evaluated using the responses for each task separately to create an even stronger competitor for the toolbox model. Next, we report the details of our approach, specifically how we proceeded in two steps: First, in Study 6a, we identified the best possible single-process model that predicts a single interval; in Study 6b, we added error distributions to generate predictions that spanned several intervals.

[8] We are grateful to one of the anonymous reviewers for raising this issue.

8.1. Study 6a: Single-interval predictions

If predictions are confined to a single interval, identifying the model that maximizes the overlap performance for each task is a straightforward endeavor. When the data to be predicted are fully known, the optimal single-process model simply predicts the interval containing the largest number of observations. Given the ability to change this interval freely between tasks, the optimal single-interval model thus identified 106 intervals (one for each task) and predicted for each task that all responses fell into this interval. No model that commits to a single interval can predict better than the relative frequency of this chosen interval.

The performance of this optimized model across tasks ranged between 13.7% and 53.6%, with an average of $M = 29.4\%$. In contrast, the Five-plus toolbox achieved an average across tasks of $M = 56.0\%$. In fact, the performance of the optimal single-interval model was only slightly better than the performance of a model that simply assumes the use of the representativeness heuristic for every single participant ($M = 24.2\%$). This is not surprising, given that the interval with the maximum number of observations contained the hit rate (representativeness heuristic) in 59 of the 106 environments (and the base rate in 31 additional environments).

In summary, the Five-plus-one toolbox model outperformed any model relying on picking a single interval. Note that, because parameters were fitted to known data, there was no genuine *prediction* involved. Rather, the exercise was one of *postdiction*. This upper limit is therefore very likely an overestimate of the maximum possible performance of any truly predictive model, irrespective of the quality criterion chosen for parameter fitting or the training data used to establish weights. This single-process model cannot be outperformed by any other single-process model that makes single predictions for each task. Yet, it was clearly outperformed by the toolbox model.

8.2. Study 6b: Allowing for errors in predictions

Next, we broadened the scope to single-process models that will permit for some degree of deviation from target intervals if it improves their performance. When introducing single-process models at the beginning of this article, we included an error term in their equations. Adding errors to their predictions in previous simulations would have increased their RMSE and would not have improved their percentage of correct predictions beyond accidental hits. When using the overlap criterion, the error modeling might prove advantageous, as it can change single-interval predictions to predictions spanning several contiguous intervals. In Study 6b, we considered single-process models that estimated error parameters to extend their predictions across intervals. In all cases, $\epsilon = 0$ was a possible value for the error term, which would result in single-interval predictions.

8.2.1. Models

As in Study 6a, we evaluated a model's quality by measuring the overlap between model estimates and actual data for each of the 106 tasks. In Study 6b, we compared four models: two postdiction models and two prediction models. In analogy to the optimal single-interval model in Study 6a, we created an optimal single-process model (OSP), which had access to the task data and chose a combination of prediction value and error parameter to maximize the overlap (see Supplementary Material, Section 6.3, for methodological details). Adding a normally distributed error to the log-odds of probabilities resulted in a slightly better model than did adding error to the probabilities (see Supplementary Material, Section 6.4). Again, we included the predictive Five-plus toolbox model (FPT) that estimated parameters with access to all nontarget task data as training data. We also included an optimal toolbox model (OTB). This postdiction model was limited to the same six rules as the predictive variant, but was able to adjust estimated frequencies based on the target data. As a final point of comparison, we included a predictive weighing-and-adding model with errors (WAD). This model had access to the same training data as the FPT model and estimated its four weighing parameters through regression, like the standard WA model. Normally distributed errors were then added to the task estimate, with the standard distribution chosen to maximize the overlap criterion in the training data. We evaluated these four models on each of the 106 tasks.

8.2.2. Results

Fig. 10 shows the performance comparison of the four models. The Five-plus toolbox model, with an average task overlap of $M = 56.0\%$, outperformed the predictive WAD model ($M = 18.0\%$) and, more importantly, it vastly outperformed the optimal single-process model ($M = 36.7\%$). Only the optimal toolbox model performed better, with $M = 62.3\%$. A closer look at the shape of the models' predictions paints an even clearer picture of the models' relative merits. We present a visualization of the four models' predictions for each of the 106 tasks in the Supplementary Material (Section 6.7). These overviews demonstrate that, on average, only 8.0% of the overlap performance of the optimal single-process model was based on responses in intervals that did not contain an estimate for one of the six toolbox rules. In additional simulations, we were further able to demonstrate that the success of the toolbox model does not depend on the full set of rules, nor on the specifics of parameter estimation: Even reduced models with fewer rules and the assumption of an equal distribution of responses still outperformed the OSP. We present this analysis and a range of further results for a variety of prediction and postdiction models in the Supplementary Material (Section 6.6).

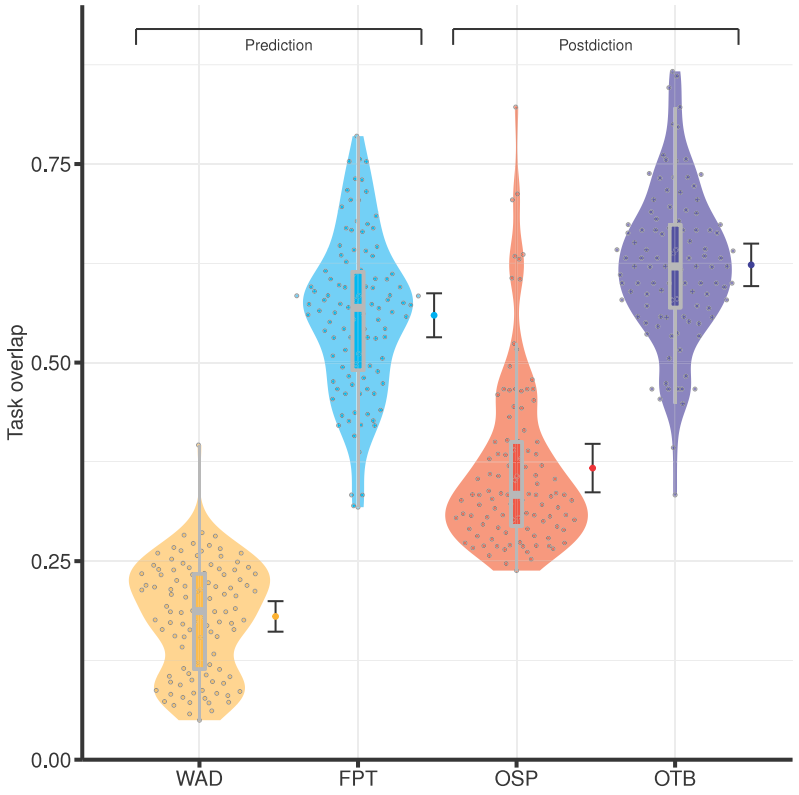


Fig. 10. Task overlap of the four models across tasks. For each model, violin plots, beeswarm plots, and boxplots show the distribution of overlap values across the 106 tasks. Dots and bars to the right of each cloud show means and the 95% confidence interval. Data is presented for the weighing-and-adding model with added simple errors (WAD), the standard Five-plus toolbox model (FPT), the optimal single-process prediction with log-odds errors (OSP), and the optimal toolbox model (OTB).

8.3. Discussion

The evidence from Study 6 is overwhelmingly in favor of the toolbox model and against single-process models of any plausible nature. We estimated the upper bound for any model predicting a single value with or without errors. Even when we granted every possible methodological advantage to single-process models (e.g., allowing them to overfit the data without penalty), they could not match the performance of the most basic toolbox model. We invite the reader to engage with the task-wise figures in the Supplementary Material (see Section 6.4). From our perspective, the demonstrated pattern across tasks is the strongest possible demonstration in favor of the toolbox model. We also show that this clear advantage does not depend on specific choices in the process of model development and testing. A defender of single-process models would have to posit that a preferred model would make different predictions for clusters of participants, or that the same process would yield responses in very specific and nonadjacent intervals through some yet-to-be-explained mechanism. Having identified a simpler, more successful model and explanation, we are inclined to discard these alternatives.

9. General discussion

We are concerned with the question of how people make an inference in tasks in which three pieces of probability information are explicitly stated: base rate, hit rate, and false alarm rate. One view in the literature assumes that people arrive at their inferences by recruiting a single process, be it the representativeness heuristic, conservatism, or the weighing-and-adding model. This single-process view assumes homogeneity in the process and inferences and, by extension, give rise to the prediction of a unimodal distribution of inference (the distribution is due to an unsystematic error component). Single-process accounts have received much attention and currency in the literature, and specifically, variants of the weighing-and-adding model are widely used, from expected utility theory to utilitarian models of moral reasoning. There is, however, an alternative view. It assumes that, across individuals, several rules will be employed when people proceed to derive an inference. This toolbox view suggests, across individuals, a multimodal distribution of inferences. Our goal was to investigate both views and find ways to examine which of the two better describes cognition in Bayesian inference tasks.

Our investigation has produced four major results. First, we found that past data reflect multimodal response distributions and thus do not support a single-process view. Second, when simulating inferences on the basis of a set of four rules, we found that a WA model can counterintuitively provide the best fit and even the best average prediction, irrespective of the fact that it entirely fails to capture the actual rule of a single simulated individual. The model's stunning ability to seemingly capture the data rests on the prediction of inferences that drift towards the middle of the probability scale. By relying on RMSE, it thereby succeeds in reducing the squared deviations between predicted and actual values better than any of the rules that actually generated the inference in the simulation. We suggest that it is this ability that explains why researchers repeatedly found and concluded that single-process explanations best capture people's Bayesian inferences.

On the basis of what is likely to be the most comprehensive analysis of individual inferences in Bayesian inference tasks, our third contribution was to observe that people use a set of five non-Bayesian rules and Bayes's rule. These jointly explain about 64% of inferences. Three of the rules consider one and only one of the three pieces of information stated (i.e., base rate, hit rate, 1 – false-alarm rate), and two other rules combine two pieces of information (i.e., hit rate – false-alarm rate, hit rate times base rate); Bayes's rule alone integrates all three pieces. This finding is based on the analysis of individual data and is supported by self-reported rule use and a response time analysis. In the latter, we found the fastest response time for those rules that rely on a single piece of information and the slowest for the two rules that require multiplication. The fourth and final contribution was to describe a procedure to deal with the seductive power of highly parameterized single-process models, namely, the predictions and test of distribution rather than point estimates. We conclude by discussing two implications of our findings.

9.1. Revisiting the cognitive limitation argument

Numerous studies have consistently shown that most people fail to compute the Bayesian posterior probability correctly from the stated pieces of information. This failure has commonly been attributed to cognitive limitations that thwart the ability for multiplication. In particular, it has been suggested that people spontaneously weigh and add probabilities instead of multiplying. Some have even argued that this process of weighing and adding is part of the "genetic endowment of Homo sapiens" (Juslin et al., 2009, p. 858). The assumption of a general-purpose additive rule is also prominent in information integration theory (Anderson, 1981). Specifically, Juslin et al. (2011) argued that (1) the cognitively limited mind is unable to perform multiplication, (2) there is clear evidence that people spontaneously weigh and add probabilities to estimate posterior probabilities,[9] and (3) the substitution of multiplication with weighing and adding need not be irrational. The reason is that in a world in which probabilities are not known but need to be estimated from limited samples of experience, both strategies converge (representing decisions from experience; see Schulze & Hertwig, 2022).

Let us have a closer look at the argument that weighing and adding replaced multiplication because it is cognitively less demanding. Consider Bayes's rule:

$$p(H|D) = \frac{p(H) \cdot p(D|H)}{p(H) \cdot p(D|H) + p(\neg H) \cdot p(D|\neg H)} \tag{7}$$

As Eq. (4) shows, Bayes's rule requires the multiplication of three terms, two of which are identical. Compare this with Eq. (3), which describes the process of the weighing-and-adding model:

$$p(H|D) = \alpha + w_b \cdot p(H) + w_h \cdot p(D|H) + w_f \cdot p(D|\neg H), \tag{3 revisited}$$

Conflicting with Juslin's argument that the human mind lacks the ability for multiplication, Eq. (3) entails three distinct multiplicative terms. These are not simple, as they involve numbers between 0 and 1. Addressing this contradiction, Juslin et al. argued that although the algebraic expression in Eq. (3) "involves multiplying cues with weights, the mental operation is not one of multiplication [...]. The process is thought to involve [...] the so-called proportional change rule (Anderson 1996, p. 60), where a currently held estimate is adjusted in proportion to the value of the information being integrated" (p. 858). We see three problems with this argument. First, the proportional change rule seems to be a redescription of what multiplication is. Second, the rule is applicable only if the weights are assumed to be positive and add up to 1, which holds solely in a special case in Eq. (3). Third, if this simpler process to implementing multiplication were in principle available, then by the same token it would also be available to implement the computations in Bayes's rule. In light of these issues, the theoretical arguments of adding and weighing and the proportional change rule hardly offer a convincing solution to the difficulty of multiplication. To avoid misunderstandings: We agree with Juslin et al. that many people are likely to seek computationally simpler processes in place of the multiplications required by Bayes's rule. Yet the empirical evidence accrued here suggests that the main path towards simplification consists in using one of

[9] One of the most cited pieces of evidence for a general-purpose additive rule in the human mind is that children spontaneously (without instruction) add and weigh the height and the width when estimating the area of a rectangle (Anderson & Cuneo, 1978). The height-and-width rule has been inferred from an analysis of variance that found the interactive term height × width to be not significant. Later it was shown that the power of the tests to detect such an interaction (assuming it exists) was typically below 10% (Gigerenzer & Murray, 2015), and that young children's ratings of area were highly unreliable. In paired comparisons, by contrast, children's judgments were highly reliable and provided evidence that no child relies on adding and weighing but that they instead use the rectangle's shape to infer its area (Gigerenzer & Richter, 1990).

the five non-Bayesian rules included in the Five-plus toolbox. All of these rules ignore some information, and only one requires multiplication.

9.2. Why does the representativeness heuristic perform so much better than conservatism?

Since the 1960s, two profoundly different views on Bayesian reasoning have dominated the discussion. Summarizing one view, Ward Edwards (1968) concluded:

> An abundance of research has shown that human beings are conservative processors of fallible information. Such experiments compare human behavior with the outputs of Bayes's theorem, the formally optimal rule about how opinions (that is, probabilities) should be revised on the basis of new information. It turns out that opinion change is very orderly, and usually proportional to numbers calculated from Bayes's theorem—but it is insufficient in amount. (p. 18)

On Edwards's view, people's intuitive reasoning is structurally Bayesian in nature, but people revise probabilities to a lesser extent than Bayes's theorem would prescribe—a behavior known as conservatism. Only a few years later, Kahneman and Tversky (1972) challenged Edwards's view and instead concluded that "in his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all" (p. 450). Based on their evidence, they suggested that people resort to the representativeness heuristic that gives too little weight to base rates, a behavior they described in terms of the base-rate fallacy.

In Study 3, we collected 30 data sets concerned with Bayesian inference tasks. With more than 4,000 participants and 106 tasks, this amounts to the largest set of responses to Bayesian tasks ever analyzed. Our analysis of these responses could be interpreted as the definitive test between the two views, with a clear winner. As Fig. 5 shows, the representativeness heuristic scored much better than the conservatism model, with a median of 24% of all inferences correctly predicted relative to the meager 0% score of the two conservatism models. Does this settle the disagreement between Edwards and Kahneman and Tversky? For an important reason, we do not think so. In a recent analysis, Lejarraga and Hertwig (2021) demonstrated that these authors arrived at their conclusions using widely different experimental methods: Edwards and colleagues employed a largely experienced-based protocol in which learning was often required (with, on average, feedback, practice trials, physical instantiations of the stimuli, and repeated trials and measurements), whereas Kahneman and Tversky turned to a largely description-based protocol offering few, if any, opportunities for learning (with, on average, no feedback, no practice trials, text-based stimuli, and fewer trials). The 30 data sets that we analyzed were all collected after Kahneman and Tversky's research program ushered in this new experimental protocol, which, as Lejarraga and Hertwig observed, became the default method. With no exception, the data sets we analyzed included description-based Bayesian inference tasks. The deck was stacked against the experiential protocol and against the way that Edwards and colleagues tested Bayesian reasoning. Therefore, we cannot settle the question of which of these views is more accurate here. Indeed, if anything, our results suggest that no single process, whether representativeness or conservatism, can sufficiently explain the heterogeneity of responses. One open question is whether, and to what extent, process heterogeneity would also be found in studies employing the experiential paradigm.

9.3. The risk of making inferences from the aggregate to the individual

From a methodological point of view, the contribution of our analyses is to highlight a time-honored lesson in psychology (see Estes, 1956): If there is heterogeneity in the cognitive processes, then inference from aggregate analysis to the individual can be highly misleading. Such inferences risk landing, metaphorically speaking, in the middle of nowhere. A second insight is that the weighing-and-adding model, equipped with several free parameters, is able to mimic fundamentally different processes such as the base-rate rule and the representativeness heuristic. Whereas the former ignores all new information, the latter ignores all old information (see Pachur et al., 2017). A final insight is that the remedy to this predicament does not necessarily consist in replacing fitting by prediction (i.e., the cross-validation approach); the seductive power of aggregation is still intact when using RSME to evaluate prediction. Instead, we propose a five-step research process to understand and model process heterogeneity.

Acknowledgments

We thank Ulrich Hoffrage and Michelle McDowell for many constructive discussions concerning the analyses and simulations reported in this article. We are grateful to Rona Unrau and Deborah Ain for editing the manuscript, and Katarzyna Dudzikowska for support in preparing the Supplementary Material. Finally, we are grateful to the authors of past articles, who without any exception, responded to our inquiries about their studies and agreed, if possible, to share their data either in repositories or through personal emails: Karin Binder, Gary Brase, Gretchen Chapman, Andrew Cohen, Thomas Dohmen, Mirta Galesic, Sebastian Hafenbrädl, Brett Hayes, Ulrich Hoffrage, Peter Juslin, Barbara Mellers, Simon McNair, Stefania Pighin, and Gaeëlle Vallée-Tourangeau.

Table B.1
Sample descriptions for the empirical datasets used in the simulations. Each sample is described by the source (or Study number for our samples), the sampled population, number of participants, the maximum number of responses per participant, the total number of responses, and the number of different tasks.

ID	Source	Sample	N	$\frac{\text{Max responses}}{\text{Participant}}$	Responses	Tasks
1	Gigerenzer and Hoffrage (1995)	Students	30	15	433	15
2	Hoffrage and Gigerenzer (1998)	Physicians	48	2	96	4
3	Hoffrage et al. (2000)	Students/Prof.	66	1	66	4
4	Juslin et al. (2011)	Students	15	18	270	18
5	Hoffrage et al. (2015)	EMBA	15	2	30	4
6	Hoffrage et al. (2015)	EMBA	14	2	28	4
7	Hoffrage et al. (2015)	EMBA	10	2	20	4
8	Hoffrage et al. (2015)	EMBA	24	2	48	4
9	Hoffrage et al. (2015)	Managers	14	2	28	4
10	Hoffrage et al. (2015)	Managers	17	2	34	4
11	Hoffrage et al. (2015)	Students	37	2	74	4
12	Hoffrage et al. (2015)	Students	21	2	42	4
13	Hoffrage et al. (2015)	Students	25	2	50	4
14	Study 4a	MTurk	545	4	2180	4
15	Study 4b	MTurk	318	2	638	4
16	Cohen and Staub (2015)	MTurk	95	36	3420	36
17	Chapman and Liu (2009)	Students	345	1	345	2
18	Galesic et al. (2009)	Younger adults	59	2	118	2
19	Galesic et al. (2009)	Older adults	23	2	46	2
20	Pighin et al. (2016)	MTurk	126	1	126	3
21	Vallée-Tourangeau et al. (2015)	Volunteers	45	3	132	3
22	Nadanovsky et al. (2018)	Dentists	52	1	52	1
23	Bar-Hillel (1980)	Univ. applic.	52	1	52	1
24	Birnbaum and Mellers (1983)	Students	65	1	65	1
25	Binder et al. (2015)	Students (16–18)	78	1	78	2
26	Weber et al. (2018)	Students	78	1	78	2
27	Hayes et al. (2018), Exp. 1	MTurk	44	2	88	2
28	Hayes et al. (2018), Exp. 2	MTurk	55	1	55	1
29	Dohmen et al. (2009)	Representat.	859	1	859	1
30	Study 4c	MTurk	1013	1	1013	8

Appendix A. Texts of exemplary bayesian inference tasks

A.1. Cab task (Bar-Hillel, 1980; originally in Kahneman & Tversky, 1972)

Two cab companies operate in a given city, the Blue and the Green (according to the color of cab they run). Eighty-five percent of the cabs in the city are Blue, and the remaining 15% are Green.

A cab was involved in a hit-and-run accident at night.

A witness later identified the cab as a Green cab.

The court tested the witness' ability to distinguish between Blue and Green cabs under nighttime visibility conditions. It found that the witness was able to identify each color correctly about 80% of the time, but confused it with the other color about 20% of the time.

What do you think are the chances that the errant cab was indeed Green, as the witness claimed? (pp. 211–212)

A.2. Weather forecast (Dohmen et al., 2009)

Imagine you are on vacation in an area where the weather is mostly sunny and you ask yourself how tomorrow's weather will be. Suppose that, in the area you are in, on average 90 out of 100 days are sunny, while it rains on 10 out of 100 days. The weather forecast for tomorrow predicts rain. On average, the weather forecast is correct on 80 out of 100 days. What do you think is the probability, in percent, that it is going to rain tomorrow? (p. 1).

A.3. Quality control (Hoffrage et al., 2015)

The Skiwell Manufacturing Company gets material from two suppliers. Supplier A's materials make up for 30% of what is used, with supplier B providing the rest. Past records indicate that 15% of supplier A's materials are defective and 10% of B's materials are defective. Since it is impossible to tell which supplier the material came from once they are in the inventory, the manager wants to know: What is the probability that material comes from supplier A given that it has been identified as defective? (p. 1).

A.4. Caries detection (Nadanovsky et al., 2018)

Imagine a population survey for interproximal caries detection using bite-wing radiographs. The information below refers to asymptomatic adults who took part in this survey. The probability that 1 of these adults has interproximal caries requiring restorative treatment, confirmed by tooth separation and direct visual and tactile examination, was 8%. The probability of having a positive bite-wing radiograph among adults who have a confirmed interproximal caries is 90%. The probability of having a positive bite-wing radiograph among adults without interproximal caries is 8%.

Question: Imagine an adult who had a positive bite-wing radiograph in the survey. What is the probability that he actually has interproximal caries? (p. 20).

Appendix B. List of datasets and tasks

See Table B.1.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cogpsych.2023.101564.

References

Anderson, N. H. (1981). *Foundations of information integration theory*. Academic Press.

Anderson, N. H., & Cuneo, D. O. (1978). The height + width rule in children's judgments of quantity. *Journal of Experimental Psychology: General*, *107*(4), 335–378. http://dx.doi.org/10.1037/0096-3445.107.4.335.

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

Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information–an empirical study on tree diagrams and 2 × 2 tables. *Frontiers in Psychology*, *6*, 1186. http://dx.doi.org/10.3389/fpsyg.2015.01186.

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. http://dx.doi.org/10.1037/0022-3514.45.4.792.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217. http://dx.doi.org/10.1037/h0047470.

Burleigh, T., Kennedy, R., & Clifford, S. (2018). How to screen out VPS and international respondents using Qualtrics: A protocol (October 12, 2018). http://dx.doi.org/10.2139/ssrn.3265459, SSRN.

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making*, *4*(1), 34–40. http://dx.doi.org/10.1017/S1930297500000681.

Cohen, A. L., & Staub, A. (2015). Within-subject consistency and between-subject variability in Bayesian reasoning strategies. *Cognitive Psychology*, *81*, 26–47. http://dx.doi.org/10.1016/j.cogpsych.2015.08.001.

Dohmen, T., Falk, A., Huffman, D., Felix, M., & Sunde, U. (2009). *The Non-Use of Bayes Rule: Representative Evidence on Bounded Rationality*. Technical Report, Maastricht University, Research Centre for Education and the Labour Market, ROA-RM-2009/1, February 2009.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). John Wiley & Sons.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: verbal reports as data*. MIT Press.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140. http://dx.doi.org/10.1037/h0045156.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. http://dx.doi.org/10.1257/089533005775196732.

Galesic, M., Gigerenzer, G., & Straubinger, N. (2009). Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Medical Decision Making*, *29*(3), 368–371. http://dx.doi.org/10.1177/0272989X08329463.

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

Gigerenzer, G., & Murray, D. J. (2015). *Cognition as intuitive statistics*. Psychology Press.

Gigerenzer, G., & Richter, H. R. (1990). Context effects and their interaction with development: Area judgments. *Cognitive Development*, *5*(3), 235–264. http://dx.doi.org/10.1016/0885-2014(90)90017-N.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*(1), 75–90. http://dx.doi.org/10.1037/0033-295X.109.1.75.

Hayes, B. K., Ngo, J., Hawkins, G. E., & Newell, B. R. (2018). Causal explanation improves judgment under uncertainty, but rarely in a Bayesian way. *Memory & Cognition*, *46*(1), 112–131. http://dx.doi.org/10.3758/s13421-017-0750-z.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*(5), 538–540. http://dx.doi.org/10.1097/00001888-199805000-00024.

Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in Psychology*, *6*, 642. http://dx.doi.org/10.3389/fpsyg.2015.00642.

Hoffrage, U., Lindsey, S., & Hertwig, R. (2000). Communicating statistical information. *Science*, *290*(5500), 2261–2262. http://dx.doi.org/10.1126/science.290.5500.2261.

Horn, S. S., Ruggeri, A., & Pachur, T. (2016). The development of adaptive decision making: Recognition-based inference in children and adolescents. *Developmental Psychology*, *52*(9), 1470–1485. http://dx.doi.org/10.1037/dev0000181.

Johnson, E. J., Payne, J. W., Bettman, J. R., & Schkade, D. A. (1989). *Monitoring information processing and decisions: The Mouselab system. Tech. rep.*, Duke University, Durham NC, Center for Decision Studies.

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*(4), 856–874. http://dx.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. *Cognition*, *120*(2), 248–267. http://dx.doi.org/10.1016/j.cognition.2011.05.004.

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

Laplace, P. S. (1814/1951). *A philosophical essay on probabilities*. Dover, Original work published in 1814.
Lejarraga, T., & Hertwig, R. (2021). How experimental methods shaped views on human competence and rationality. *Psychological Bulletin*, *147*(6), 535–564. http://dx.doi.org/10.1037/bul0000324.
Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior and Human Decision Processes*, *82*(2), 217–236. http://dx.doi.org/10.1006/obhd.2000.2895.
McKenzie, C. R. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology*, *26*(3), 209–239. http://dx.doi.org/10.1006/cogp.1994.1007.
Mellers, B. A., & McGraw, A. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychological Review*, *106*(2), 417–424. http://dx.doi.org/10.1037/0033-295X.106.2.417.
Nadanovsky, P., dos Santos, A. P. P., Lira-Junior, R., & de Oliveira, B. H. (2018). Clinical accuracy data presented as natural frequencies improve dentists' caries diagnostic inference: Evidence from a randomized controlled trial. *The Journal of the American Dental Association*, *149*(1), 18–24. http://dx.doi.org/10.1016/j.adaj.2017.08.006.
Pachur, T., Suter, R. S., & Hertwig, R. (2017). How the twain can meet: Prospect theory and models of heuristics in risky choice. *Cognitive Psychology*, *93*, 44–73. http://dx.doi.org/10.1016/j.cogpsych.2017.01.001.
Piaget, J., & Inhelder, B. (1951/1975). *The origin of the idea of chance in children*. Norton, Original work published in 1951.
Pighin, S., Gonzalez, M., Savadori, L., & Girotto, V. (2016). Natural frequencies do not foster public understanding of medical test results. *Medical Decision Making*, *36*(6), 686–691. http://dx.doi.org/10.1177/0272989X16640785.
Schulte-Mecklenbeck, M., Kühberger, A., & Johnson, J. G. (2011). *A handbook of process tracing methods for decision research: a critical review and user's guide*. Psychology Press.
Schulze, C., & Hertwig, R. (2022). Experiencing statistical information improves children's and adults' inferences. *Psychonomic Bulletin & Review*, *29*, 2302—2313. http://dx.doi.org/10.3758/s13423-022-02075-3.
Shepard, R. N. (1990). *Mind sights: original visual illusions, ambiguities, and other anomalies, with a commentary on the play of mind in perception and art*. WH Freeman/Times Books/Henry Holt & Co.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. http://dx.doi.org/10.1037/0033-295X.90.4.293.
Vallée-Tourangeau, G., Abadie, M., & Vallée-Tourangeau, F. (2015). Interactivity fosters Bayesian reasoning without instruction. *Journal of Experimental Psychology: General*, *144*(3), 581–603. http://dx.doi.org/10.1037/a0039161.
Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology, Vol. 1* (pp. 135–151). Penguin.
Weber, P., Binder, K., & Krauss, S. (2018). Why can only 24% solve Bayesian reasoning problems in natural frequencies: Frequency phobia in spite of probability blindness. *Frontiers in Psychology*, *9*, 1833. http://dx.doi.org/10.3389/fpsyg.2018.01833.
Woike, J. K. (2019). Upon repeated reflection: Consequences of frequent exposure to the cognitive reflection test for Mechanical Turk participants. *Frontiers in Psychology*, *10*, 2646. http://dx.doi.org/10.3389/fpsyg.2019.02646.
Woike, J. K., Hoffrage, U., & Martignon, L. (2017). Integrating and testing natural frequencies, naïve Bayes, and fast-and-frugal trees. *Decision*, *4*(4), 234–260. http://dx.doi.org/10.1037/dec0000086.
Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition*, *98*(3), 287–308. http://dx.doi.org/10.1016/j.cognition.2004.12.003.