Judgment and Decision Making, Vol. 14, No. 1, January 2019, pp. 1–10

Belief bias and representation in assessing the Bayesian rationality of others

Richard B. Anderson* Laura Marie Leventhal† Don C. Zhang‡ Daniel Fasko, Jr.§
Zachariah Basehore¶ Christopher Gamsby¶ Jared Branch¶ Timothy Patrick¶

Abstract

People often assess the reasonableness of another person's judgments. When doing so, the evaluator should set aside knowledge that would not have been available to the evaluatee to assess whether the evaluatee made a reasonable decision, given the available information. But under what circumstances does the evaluator set aside information? On the one hand, if the evaluator fails to set aside prior information, not available to the evaluatee, they exhibit belief bias. But on the other hand, when Bayesian inference is called for, the evaluator should generally incorporate prior knowledge about relevant probabilities in decision making. The present research integrated these two perspectives in two experiments. Participants were asked to take the perspective of a fictitious evaluatee and to evaluate the reasonableness of the evaluatee's decision. The participant was privy to information that the fictitious evaluatee did not have. Specifically, the participant knew whether the evaluatee's decision judgment was factually correct. Participants' judgments were biased (Experiments 1 and 2) by the factuality of the conclusion as they assessed the evaluatee's reasonableness. We also found that the format of information presentation (Experiment 2) influenced the degree to which participants' reasonableness ratings were responsive to the evaluatee's Bayesian rationality. Specifically, responsivity was greater when the information was presented in an icon-based, graphical, natural-frequency format than when presented in either a numerical natural-frequency format or a probability format. We interpreted the effects of format to suggest that graphical presentation can help organize information into nested sets, which in turn enhances Bayesian rationality.

Keywords: belief bias, reasoning, Bayesian inference, rationality, counterfactual

1 Introduction

People often need to assess the reasonableness of another person's judgments. For example, one might assess whether a physician, a set of jurors, a referee for a journal, or a political leader has given proper consideration to all data that were available at the time of a crucial decision. In such situations, the evaluator should set aside any knowledge he or she may have that was not available to the evaluatee (the person being evaluated), and assess whether the evaluatee made a reasonable decision. Thus, while there is a general Bayesian requirement (Bayes, 1763; Eddy, 1982; Peterson & Miller, 1965) judgments be made on the basis of complete rather than incomplete prior knowledge, the set of relevant information should include only that which the evaluatee is in a position to know. Though past research has examined the improper neglect of prior information in judgments (e.g., base-rate neglect), no research has explored the degree to which people correctly ignore prior information when judging the rationality of others. In the present paper, we explore the factors that are in play as the evaluator considers the decisions of the evaluatee. Three phenomena of interest are (1) *belief bias*, (2) the potential for *counterfactuality* to *enhance discriminability*, and (3) the *sub-optimal combination of relevant probability information* (for example, base-rate neglect).

1.1 Failure to consider all relevant probabilities: Base rate neglect

In Bayesian judgment tasks, people often fail to adequately consider prior probabilities (i.e., base-rates — Bar-Hillel, 1980; Eddy, 1982; Sloman, Over, Slovak & Stibel, 2003; Tversky & Kahneman, 1982). For example, suppose the outcome of a particular medical diagnostic test, with a known accuracy rate, indicates that a patient has Disease A. Under such conditions, research participants do consider the test's accuracy, but they do not give adequate consideration to the prior probability of having the disease. That is they neglect to properly consider what the probability of having the disease would be, prior to knowing the test's outcome. The value of

*Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403. Email: randers@bgsu.edu
†Computer Science Department, The University of Findlay.
‡Louisiana State University, Department of Psychology.
§School of Educational Foundations, Leadership, and Policy, Bowling Green State University.
¶Department of Psychology, Bowling Green State University.

https://doi.org/10.1017/S1930297500002862 Published online by Cambridge University Press

this prior probability depends on the disease's base rate — i.e., its prevalence within the population.

Because base-rate neglect (or generally, prior-probability neglect) is well-established within the literature, one would expect it to replicate in situations, such as the present studies, wherein an evaluator assesses the reasonableness of an evaluatee's probability judgments. Such replication is not guaranteed, however, since a number of factors, to be reviewed in a later section, can influence the magnitude of base-rate neglect.

1.2 Knowing that someone's prediction has been falsified by an outcome

Will an evaluator's judgment of reasonableness be biased by knowledge that the evaluatee has made a counterfactual judgment? A number of studies have examined situations in which an evaluator is aware of a valenced outcome (one that is favorable or unfavorable) that could not have been known by evaluate, at the time the evaluatee made his or her judgment (Baron & Hershey, 1988; Savani & King, 2015; Sezer, Zhang, Gino & Bazerman, 2016). Succinctly, "People view the same decision as better when it is followed by a positive outcome than by a negative outcome, a phenomenon called the outcome bias" (Sezer et al., 2016, p. 1). For example, a surgeon's decision to operate seems more reasonable when the evaluator knows the surgery succeeded than when it is known to have failed (Baron & Hershey, 1988).

However, it is not known whether mere knowledge that an evalatee's judgment has turned out to be correct or incorrect (as opposed to knowing the positivity/negativity of a judgment's impact) is enough to bias an evaluator. Suggestive evidence comes from research on people's ability to reason logically about syllogisms: In a phenomenon termed *belief bias*, people judge an argument to be less valid when they have a prior belief in the falsity of the argument's conclusion (e.g., Dube, Rotello & Heit, 2010; Evans, Handley & Harper, 2001; Newstead, Pollard, Evans & Allen, 1992; Trippas, Handley & Verde, 2013). It may be that an analogous effect occurs with respect to judgments that involve probabilities. That is, an evaluator's knowledge that the evaluatee's conclusion, prediction, or diagnosis is counterfactual may bias the evaluator to think the evaluatee's judgment was unreasonable.

Distinct from outcome bias is the phenomenon wherein a person's knowledge of the occurrence of a particular event can cause that person to either overestimate the extent to which he or she could have successfully predicted the event, or to falsely remember the he or she made a correct prediction (this is "hindsight bias;" see Pohl, 2017, for a review). However, it is not known whether knowledge of an event's occurrence has the additional effect of making those who lack such knowledge — through no fault of their own — appear unreasonable.

1.3 Summary, and overview of the present studies

On each trial of the two experiments in the present paper, a fictitious evaluatee made a diagnosis that was either *Bayes consistent* (i.e., that took base rates into account), or that was *Bayes inconsistent* (i.e., that neglected the base rates). Additionally, the diagnosis was either *factual* (it was a correct match to reality) or *counterfactual* (it was incorrect). We sought to address the following questions raised in the literature review.

Are evaluators' judgments of reasonableness responsive to the evaluatee's use or non-use of base rates? We attempted to answer this question by manipulating the Bayes consistency (consistent or inconsistent) of the evaluatee's judgment, and by measuring the evaluator's rating of the reasonableness of the evaluatee's judgment.

Are evaluators' judgments of reasonableness biased in the sense that they are responsive to the factuality (factual vs. counterfactual) of the evaluatee's judgment? To answer this question, we manipulated the factuality of an evaluatee's judgment, and we measured the evaluators' rating of the evaluatee's reasonableness.

When an evaluator knows that the evaluatee has made a counterfactual diagnosis, does such knowledge help the evaluator discriminate objectively reasonable predictions (on the part of the evaluatee) from unreasonable ones? We addressed this question by assessing whether the evaluators' ratings of reasonableness were responsive to the evaluatees' Bayes consistency, and (b) whether there was an interaction with factuality such that responsiveness was greater when the evaluatee's diagnosis was counterfactual than when it was factual.

We also investigated the potential effects of information format (probability, versus numerical natural frequency, versus graphical natural frequency), but we defer discussion of that factor until later in the paper, since it is only relevant to Experiment 2.

2 Experiment 1

The participant played the role of an evaluator, assessing the reasonableness of the judgment of a fictitious physician who has diagnosed a woman as being pregnant with a single fetus or with twins. The diagnosis was either Bayes consistent (in that it was consistent with the provided base rates for single-fetus versus twin pregnancies) or Bayes inconsistent, and was either factual (i.e., a correct assessment of the woman's actual pregnancy status), or counterfactual. We expected these independent variables to have main effects as well as interactive effects.

https://doi.org/10.1017/S1930297500002862 Published online by Cambridge University Press

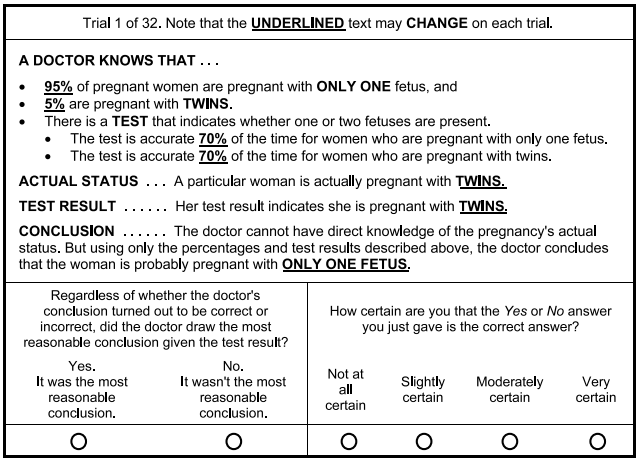


FIGURE 1: Illustration of the procedure for a single trial. In this example, the diagnosis is *counterfactual* because it does not match the pregnancy's actual status. Additionally, the diagnosis of "single fetus" is Bayes-consistent because it is consistent with the expressed base rates (which are so extreme that they overwhelm the test's diagnostic accuracy). We combined the results of the yes/no and the certainty rating to compute a reasonableness rating. The independent variables (IV) are the Bayes-consistency and factuality of the diagnosis. Note that the information about actual status is irrelevant to the judgment of reasonableness. In this example, the probability that the test result is correct, given the evidence, is 0.11.

2.1 Method

2.1.1 Participants

The participants (N = 98; 67 women) were undergraduate psychology students at Bowling Green State University. They ranged in age from 18 to 27 (mean of 19.2). They volunteered by selecting the study from a list of studies advertised on a web page designed for recruiting Bowling Green State University students. The experiment was conducted via the internet, and was implemented using Qualtrics. Participants received credit toward their course requirements in exchange for their participation.

2.1.2 Design

The experiment was a 2 by 2 within-participant factorial. One independent variable was the factuality/counterfactuality of the physician's conclusion. (A factual conclusion was a diagnosis that matched the woman's actual status; otherwise the conclusion was counterfactual.) A second independent variable was the *Bayes-consistency* of the physician's diagnosis: Given the pregnancy test result, *together with the base rates* for the two kinds of pregnancy (single or twin fetuses), the conclusion (the diagnosis) was either Bayes-consistent (i.e., consistent with the conclusion prescribed by Bayes theorem), or it was Bayes-inconsistent. Thus, we manipulated Bayes-consistency independently of factuality/counterfactuality. To avoid exact repetition of trial content, the accuracy-percentage for the diagnostic test also varied slightly across trials (70% or 75%), as did the base rates. For single-fetus and twin pregnancies, respectively, the base rates were either 90% and 10%, or 95% and 5%). All variations in trial content, and all manipulations, were within-participant.

In each trial, the participant's response was a yes/no judgment of the reasonableness of a medical diagnosis, along with a confidence rating for the yes/no judgment. Figure 1 illustrates a single trial. Table 1 indicates the content of the 32 trials, which were randomly ordered for each participant.

2.1.3 Procedure

Each participant performed 32 trials of a medical judgment task (Figure 1) wherein participants rated the reasonableness of a physician who has access to base rates, along with a diagnostic test, to assess whether a pregnant woman is carrying a single fetus or twins. The order of the trials was randomized separately for each participant. As indicated in Figure 1, each trial included information about base rates, the expected accuracy of the test, and the factuality of the physician's diagnosis.

2.2 Results

Each assessment of reasonableness was converted to an eight-point scale ranging from −4 (*very certain "no"*) to +4 (*very certain "yes"*), by combining the yes/no and rating scale assessments shown in Figure 1.

To assess whether the data might be compatible with a signal detection model, we examined ROC curves defined as the hit rate as a function of the false alarm rate (see Macmillan & Creelman, 2005, for a description of signal detection modeling). However, we found that the curves were asymmetrical, thus violating the signal detection analysis prerequisite that each curve be symmetrical. Consequently, we do not report signal detection analyses in the present paper.

As shown in Figure 2, an analysis of variance (ANOVA) produced an expected main effect of the Bayes-consistency of the of the diagnosis (made by the fictitious physician) on participants' reasonableness ratings [$F(1, 97) = 30.53$, $p < .001$, $\eta_p^2 = .24$]. The figure also shows a main effect of factuality on participants' reasonableness ratings [$F(1, 97) = 73.34$, $p < .001$, $\eta_p^2 = .43$]. Additionally there was an interactive effect of Bayes-consistency and factuality on participants' reasonableness ratings: The effect of Bayes consistency on the mean rating was greater in the counterfactual condition than in the factual condition [$F(1, 97) = 7.37$, $p = .008$, η_p^2

https://doi.org/10.1017/S1930297500002862 Published online by Cambridge University Press

TABLE 1: Stimulus-Set Structure for Experiment 1.

Trial ID	Base Rates (Single Fetus; Twins)	Test Accuracy (Single Fetus; Twins)	Test Result	Conclusion (diagnosis)	Actual Status	Bayes-Consistency of Diagnosis (IV)	Factuality of Diagnosis (IV)
1	90%; 10%	70%; 70%	O	O	O	consistent	factual
2	90%; 10%	70%; 70%	T	O	O	consistent	factual
3	90%; 10%	70%; 70%	O	T	T	inconsistent	factual
4	90%; 10%	70%; 70%	T	T	T	inconsistent	factual
5	90%; 10%	70%; 70%	O	O	T	consistent	counterfactual
6	90%; 10%	70%; 70%	T	O	T	consistent	counterfactual
7	90%; 10%	70%; 70%	O	T	O	inconsistent	counterfactual
8	90%; 10%	70%; 70%	T	T	O	inconsistent	counterfactual
...							

Note. O = only one fetus, T = twins. IV indicates that the variable is an independent variable in the design. The stimuli were designed so that a Bayes-consistent diagnosis was always "only one fetus" (given the extreme base rates favoring "only one fetus"). The table includes only 8 of the 32 stimulus configurations. The remaining 24 configurations follow the same pattern except that the Test Accuracy percentages were sometimes 75 and 75 instead of 70 and 70, and the base rate percentages were sometimes 95 and 5 instead of 90 and 10. The design was evenly counterbalanced across the aforementioned variable levels.

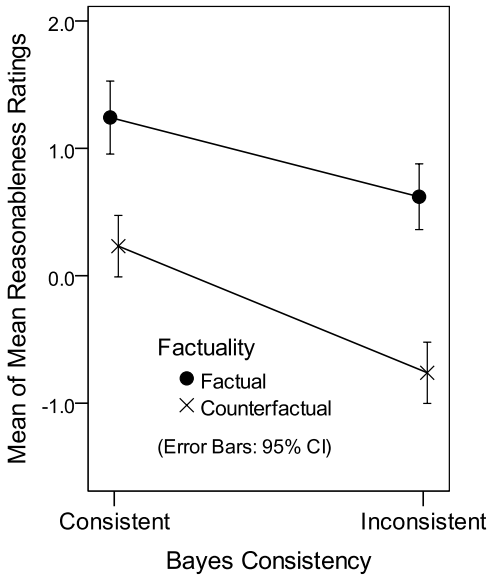


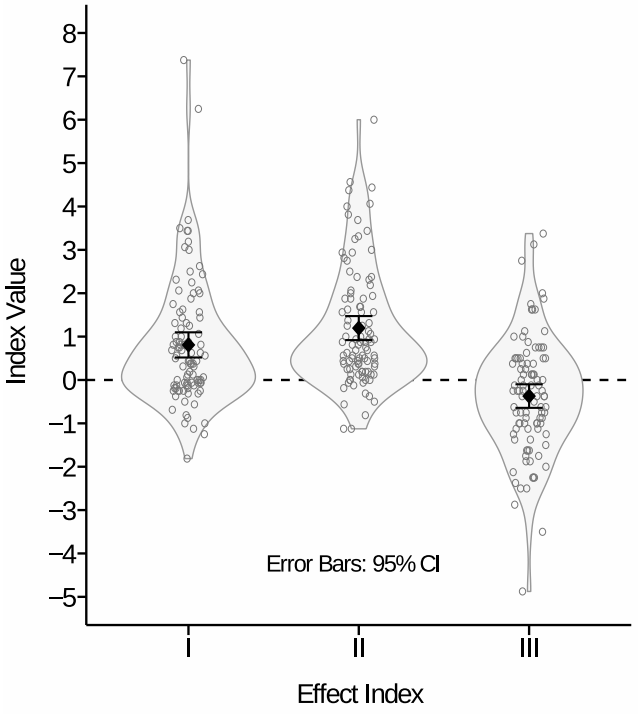
FIGURE 2: Experiment 1. Reasonableness ratings as a function of the Bayes-consistency and factuality of a physician's conclusion. N = 98.

= .07]. Because the interaction did not entail a change in the effect's direction it is possible (though not guaranteed) that it occurred as an artifact of the particular scaling of the dependent variable (see Wagenmakers, Krypotos, Criss & Iverson, 2012, for a discussion of "removable" interactions).

We also analyzed participants' binary assessments of reasonableness (the yes/no judgments concerning whether the doctor drew the most reasonable conclusion), since these assessments were not weighted by the confidence judgments and were therefore less complicated, conceptually, than the rating scale data. Each binary assessment of "reasonable" was scored as 1; each assessment of "unreasonable" was scored as −1. For each participant and in each condition, we calculated the mean value for the binary assessment. The data pattern was similar to that obtained from the rating scale data. There was a significant main effect of Bayes consistency, a significant main effect of factuality, and a significant interaction, ps < .005. For the factual conditions: M = .33 (95% CI [.24, .42]) in the Bayes-consistent condition, and M = .17 (95% CI [.09, .25]) in the Bayes-inconsistent condition. For the counterfactual conditions: M = .07 (95% CI [−.02, .16]) for the Bayes-consistent condition, and M = −.23 (95% CI [−.32, −.15]) for the Bayes-inconsistent condition.

Figure 3 shows two other indices of individual participants' performance. The first was an index of the main effect of the Bayes consistency of the stimulus problem on the reasonableness rating. We calculated this by subtracting each participant's mean reasonableness rating in the Bayes inconsistent condition from the participant's mean reasonableness rating in the Bayes consistent condition. Second was an index of the main effect of counterfactuality on the reasonableness rating. We calculated this by subtracting the participant's mean reasonableness rating in the counterfactual condition from the participant's mean reasonableness rating in the factual condition.

https://doi.org/10.1017/S1930297500002862 Published online by Cambridge University Press



I. Effect of Bayes Consistency on Reasonableness Rating
II. Effect of Factuality on Reasonableness Rating
III. I × II Interaction

FIGURE 3: Effect-indices computed for each participant in Experiment 1.

2.3 Discussion

The present main effect of factuality on ratings of reasonableness — that is, the pattern of higher ratings for factual than for counterfactual diagnoses (Figure 2) — demonstrates that the belief bias, defined as the failure to exclude irrelevant information (e.g., Evans et al., 2001), can occur within the context of Bayesian judgment. This is because any effect of factuality (on any of our dependent measures) indicated participants' tendency to consider factuality (i.e., to let factuality impact their judgment) when they should not. The interaction, in which the effect of Bayes-consistency on rated reasonableness was greater for counterfactual than for factual diagnoses (Figure 2) was a weaker effect, and notably, did not involve a change in the direction of the effect. Thus it is possible that the interaction is a scaling artifact (see Wagenmakers, Krypotos, Criss & Iverson, 2012). The present main effects of factuality and Bayes consistency demonstrated limited rationality in the participants. In judging the reasonableness of another (fictitious) person's judgments, participants were responsive to a relevant factor: the Bayes-consistency of a fictitious person's judgment. But they were also responsive to an irrelevant factor: the factuality of such fictitious judgments.

3 Experiment 2

Experiment 2 was an attempt to replicate the major findings of Experiment 1, which were that people's ratings of the reasonableness of a judgment were responsive both to the Bayes-consistency and the factuality of the judgment. As will be discussed in the following section, information format is known to affect performance on judgment and decision-making tasks, often for reasons that are unclear. But an important step in assessing the replicability of the effects obtained in Experiment 1 is to attempt such replications with multiple information formats.

Graphical visual aids can facilitate judgments of frequency and of probability (Garcia-Retamero & Cokely, 2011, 2013, Okan, Garcia-Retamero, Cokely & Maldonado, 2015). A particular kind of visual aid known as an icon array has been shown to reduce people's tendency to neglect the denominator portion of a relative frequency (Garcia-Retamero, Galesic & Gigerenzer, 2010). To illustrate this neglect: When told that 98 out of 3500 people who took Drug X had a fatal reaction, and 1 out of 10 who took Drug Y had a fatal reaction, people's assessments of the relative danger of the two drugs do not sufficiently consider the 3500 and the 10. However, there have not been clear, consistent findings showing that the use of icon arrays, or other kinds of graphics, reduces base rate neglect beyond that achieved by the use of a natural frequency format. For example, Sedlmeier and Gigerenzer (2001) found no advantage of graphical presentation over a natural frequency format, and in the context of a training study, Talboy and Schneider (2017) found no overall advantage of graphical over natural frequency format. Note, however, that graphical training produced greater subsequent facilitation on graphical problems than on natural frequency problems, and likewise, natural frequency training produced greater subsequent facilitation on natural frequency problems than on graphically presented problems.

Like Experiment 1, Experiment 2 required participants to make judgments about the reasonableness of the conclusions of a fictitious physician. However, Experiment 2 employed multiple stimulus-presentation formats. In one condition, the stimulus data were presented as probabilities (technically, as percentages). In a second condition the stimuli were presented in a numerical natural-frequency format (indicating the joint frequencies for the two possible test results and the two possible pregnancy statuses). A third condition employed an icon-based, graphical, natural-frequency format. Because prior research (e.g., Gigerenzer & Hoffman, 1995; Sedlmeier & Gigerenzer, 2001; Sloman et al., 2003; Talboy & Schneider, 2017) has implicated natural-frequency formats as facilitators of Bayesian inference, a reasonable expectation was that Experiment 2 would produce greater sensitivity to Bayesian rationality, and less bias in the graphical and numerical natural-frequency conditions than in the probability condition. We did not have strong expectations

https://doi.org/10.1017/S1930297500002862 Published online by Cambridge University Press

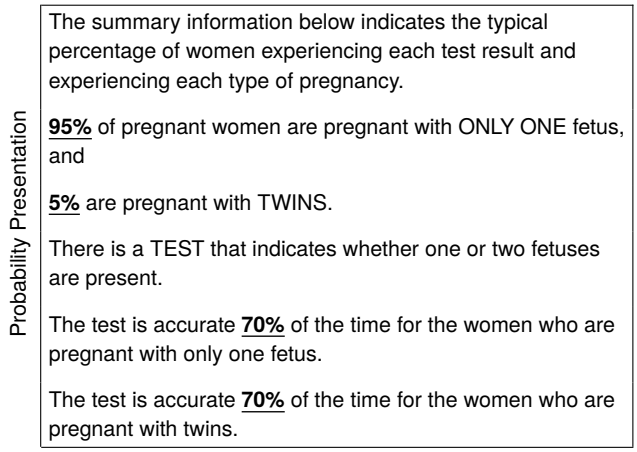


FIGURE 4: Example of information presented to participants in the *probability* condition, in Experiment 2.

concerning possible effects of icon-based, graphical presentation, since prior research did not provide a strong basis for such expectations.

3.1 Method

3.1.1 Participants

One hundred forty-one participants (94 women) were recruited through Amazon's Mechanical Turk. Each was paid $2.00. They participated via the internet, using a web browser. Their ages ranged from 18 to 29 with a mean of 18.9.

3.1.2 Design

The experimental design was like that of Experiment 1 except that Experiment 2 included three stimulus-data formats rather than just one. The three formats were: probability (Figure 4), graphical natural-frequency (Figure 5), and numerical natural-frequency (also Figure 5). Thus, the design was a 3 by 2 by 2 factorial: *Format* (probability, numerical natural-frequency, or graphical natural-frequency), varied between subjects, *factuality* (factual or counterfactual), and *Bayes-consistency* (consistent or inconsistent) varied within subjects. As in Experiment 1, the dependent variable was the rating of reasonableness.

3.1.3 Procedure

Each participant was randomly assigned to one of three data format conditions (probability, numerical natural-frequency, or graphical natural-frequency) illustrated in Figures 6 and 7. Positioned below the data was information about the "actual status," "test result," and "conclusion," along with a response scale: This was the same information and scale used in Experiment 1. In addition to the 32 trial types used in Experiment 1, the present procedure included four attentiveness-test trials randomly interspersed throughout the trial sequence. (Trial order was randomized, separately for each participant.) On such trials the stimulus display excluded crucial information. Specifically, it excluded information concerning the test result, the actual status, and the physician's conclusion. In place of such information was an instruction to make a specific set of responses, such as "Please answer 'No' and 'Very Certain' "). Thus, each participant could receive an attentiveness score (0 to 4) indicating the number of correct responses to the attentiveness questions.

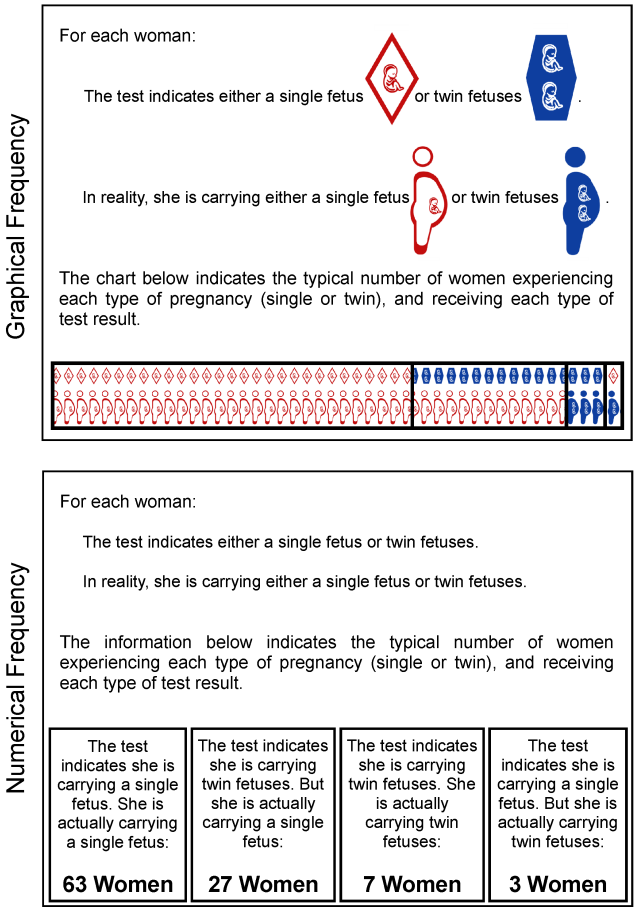


FIGURE 5: Example of information presented to participants in the *graphical natural-frequency* and *numerical natural-frequency* conditions in Experiment 2.

3.2 Results and discussion

As in Experiment 1, we conducted separate analyses for the quantitative (8-point scale) reasonableness assessment and the binary (yes/no) assessment. To facilitate visual comparison of the results for the two analyses, a "yes" on the binary scale was scored as a 4.0, and a "no" was scored as a −4.0. Figure 6 shows the data pattern and Table 2 presents the ANOVA results.

https://doi.org/10.1017/S1930297500002862 Published online by Cambridge University Press

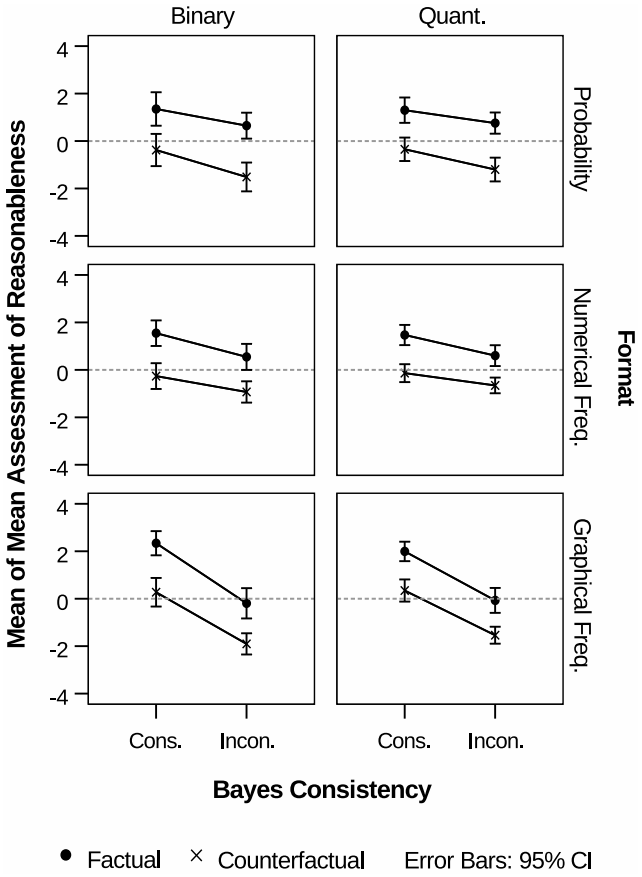


FIGURE 6: Effects of factuality, Bayes-consistency (consistent or inconsistent), and information format (probability, numeric natural-frequency, or graphical natural-frequency) on participants' assessments of reasonableness. To facilitate graphical comparisons of the data patterns for binary versus quantitative (−4 to 4) scale data, each binary response value is scored as +4 or −4).

Table 2 indicates there were significant main effects of a conclusion's Bayes-consistency, and its factuality, on the mean assessment of reasonableness (whether the assessments are measured on a binary or quantitative scale). Experiment 2 therefore replicated principal aspects of the results of Experiment 1: The present main effect of Bayes-consistency indicates that people are sensitive to a fictitious decision-makers' Bayesian rationality, and the main effect of factuality indicates a bias to consider irrelevant information — that is, to consider factuality — in assessing the reasonableness of the conclusion.

There were also two-way interactions in which the effect of Bayes consistency varied significantly across the levels of information format. Additionally, the effect of Bayes consistency was greater for the graphical natural-frequency format than for either the numeric natural-frequency format or the probability format (Table 3).

TABLE 2: Analyses of variance: quantitative and binary assessments of reasonableness of several measures.

ANOVA dependent variable	η^2_p	df	F	p
Bayes Consistency				
Quantitative Response	.243	1, 138	44.37	<.001
Binary Response	.251	1, 138	46.24	<.001
Factuality				
Quantitative Response	.418	1, 138	98.99	<.001
Binary Response	.349	1, 138	73.58	<.001
Format				
Quantitative Response	.013	2, 138	0.91	.404
Binary Response	.008	2, 138	0.53	.589
Bayes Consistency by Factuality				
Quantitative Response	.003	1, 138	0.36	.550
Binary Response	.002	1, 138	0.23	.632
Bayes Consistency by Factuality				
Quantitative Response	.097	2, 138	7.45	.001
Binary Response	.092	2, 138	7.01	.001
Bayes Consistency by Format				
Quantitative Response	.006	2, 138	0.40	.670
Binary Response	.003	2, 138	0.18	.836
Bayes Consistency by Factuality by Format				
Quantitative Response	.031	2, 138	2.18	.117
Binary Response	.028	2, 138	1.95	.146

TABLE 3: Experiment 2. Supplemental analyses assessing the interaction between format and Bayes consistency, with only two levels of format in each analysis.

ANOVA dependent variable	η^2_p	df	F	p
Graphical Frequency vs. Probability				
Quantitative Response	.073	1, 97	7.67	.007
Binary Response	.067	1, 97	6.98	.010
Graphical Frequency vs. Numerical Frequency				
Quantitative Response	.083	1, 102	9.25	.003
Binary Response	.083	1, 102	9.29	.003

The number of participants receiving an attentiveness score of 4.0, 3.0, 2.0, 1.0, or 0.0, was 75, 18, 15, 9, and 24, respectively. The large number of less-than-4.0 scores raises the question of whether the observed data patterns are evident at all levels of attentiveness. Though some of the

https://doi.org/10.1017/S1930297500002862 Published online by Cambridge University Press

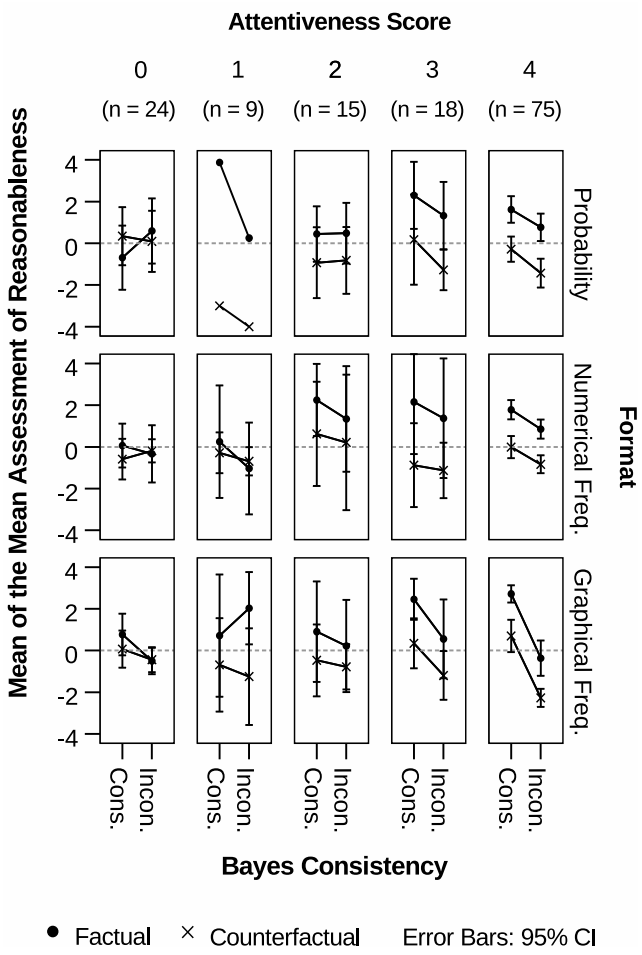
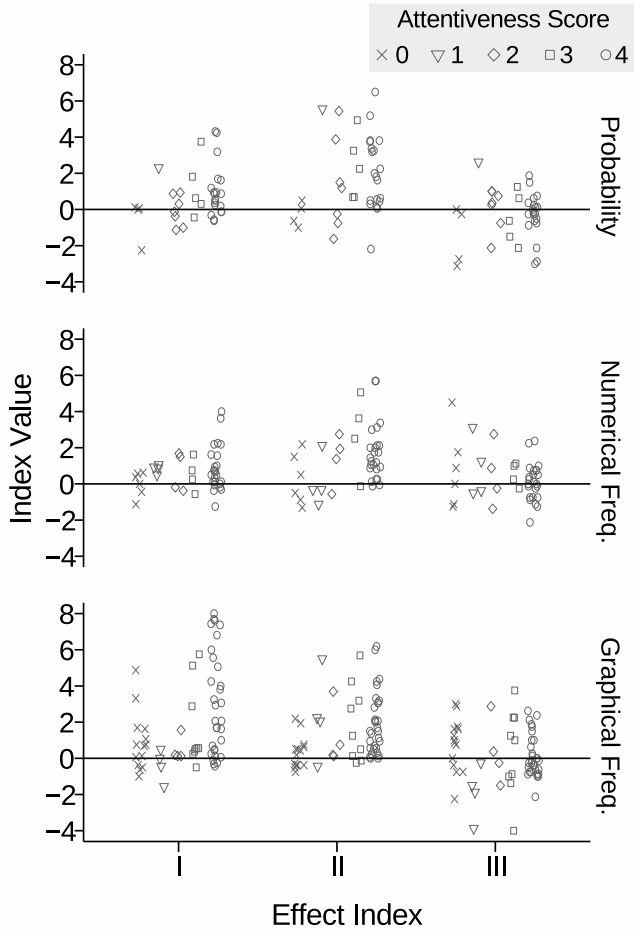


FIGURE 7: Quantitative ratings of reasonableness as functions of attentiveness score, factuality, Bayes-consistency (consistent or inconsistent), and information format (probability, numeric natural-frequency, or graphical natural-frequency).



I. Effect of Bayes Consistency on Reasonableness Rating
II. Effect of Factuality on Reasonableness Rating
III. I × II Interaction

FIGURE 8: Effect-indices as a function of attentiveness score (0 through 4) and presentation format (graphical frequency, numerical frequency, probability), computed for each participant in Experiment 2.

cells sizes in Figure 8 are very small (in some cases, there is only one observation in a cell), the data in Figure 7 suggested that the pattern of the interaction between Bayes consistency and format becomes progressively more evident as participants' attentiveness to the task increases. (And note that for participants achieving the maximum attentiveness score, the interaction was significant, $F(2, 138) = 7.45$, $p = .001$, $\eta^2_p = .097$). Additionally, Figure 8 shows the effect index values for each participant.

Taken together, the results indicate that people were most sensitive to Bayes consistency when the information was presented in a graphical natural-frequency format. This facilitative effect is consistent with previous research indicating that graphical representation can enhance probability-related judgments (e.g., Garcia-Retamero, Galesic & Gigerenzer, 2010; Sedlmeier & Gigerenzer, 2001; Okan, Garcia-Retamero, Cokely & Maldonado, 2015; Sloman, Over, Slovak & Stibel, 2003).

In contrast to Experiment 1, Experiment 2 produced no interaction between Bayes consistency and factuality, thus providing no evidence to favor a selective scrutiny mechanism — identified by Trippas et al. (2013) in the context of syllogistic reasoning, wherein counterfactuality triggers extra scrutiny, which in turn triggers enhanced discrimination between rational and irrational inferences.

4 General discussion

In the present set of studies, we used a Bayesian inference task to investigate people's perception of others' rational-

https://doi.org/10.1017/S1930297500002862 Published online by Cambridge University Press

ity. In both studies, people showed some sensitivity to the Bayes-consistency of another (fictitious) person's conclusions; the ratings of the reasonableness of such conclusions were higher for Bayes-consistent than for Bayes-inconsistent conclusions. We also found in Experiment 2 that sensitivity was enhanced by a graphical natural-frequency format, as opposed to either a numerical natural-frequency or a probability format. Specifically, the effect of the Bayes-consistency of a conclusion on the ratings of the reasonableness of that conclusion was greatest in the graphical natural-frequency condition. There was clear evidence of a graphical advantage that was not attributable to the fact that the graphical format was also a natural-frequency format.

A potential explanation for this finding is that, beyond explicating the natural frequencies, the present graphical format (see the illustration in Figure 5) served to organize the frequencies into nested sets (i.e., sets within sets; see Barbey & Sloman, 2007, for a discussion), that help the decision maker conceptualize probabilities (in the form of proportions) and natural frequencies simultaneously. The graphical bar in Figure 5 is divided into four sections, with the width of each section indicating a natural frequency. However, the same graphic is divisible into two larger sections that help explicate a pair of probabilities. There is a left section that indicates the frequency of women who are carrying a single fetus, but that also shows the proportion of those women diagnosed as carrying twins (the size relationship between the left section — which consists of the concatenation of the first and second sections — and the second section). Likewise, the left section of the graphic shows not just the frequency of women carrying twins but also the proportion of those women diagnosed as carrying a single fetus.

In summary, the present findings demonstrate the existence of belief biases in evaluating the rationality of others: There was a bias to consider information that could not have been available to the person being evaluated. The present findings also showed that the responsiveness of assessed rationality to the Bayes consistency of another person's conclusion was greater with a graphical frequency format than with either a numerical frequency or a probability format. This result was interpreted to indicate a facilitatory role for nested-set representation, and particularly for icon-based graphic representation, in Bayesian judgment.

5 References

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

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

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology, 54*, 569–579. http://dx.doi.org/10.1037/0022-3514.54.4.569.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370–418. http://dx.doi.org/10.1098/rstl.1763.0053.

Dube, C., Rotello, C. M., Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect, *Psychological Review*, *117*, 831–863. http://dx.doi.org/10.1037/a0019634.

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

Evans, J. St. B. T., Handley, S. J., & Harper, C. N. J. (2001). Necessity, possibility, and belief: A study of syllogistic reasoning. *The Quarterly Journal of Experimental Psychology*, *54A*, 935–958. http://dx.doi.org/10.1080/02724980042000417.

Garcia-Retamero, R., & Cokely, E. T. (2011). Effective communication of risks to young adults: Using message framing and visual aids to increase condom use and STD screening. *Journal of Experimental Psychology: Applied, 17*, 270–287. http://dx.doi.org/10.1037/a0023677.

Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science, 22*, 392–399. http://dx.doi.org/10.1177/0963721413491570.

Garcia-Retamero, R., Galesic, M., & Gigerenzer, G. (2010). Do icon arrays help reduce denominator neglect? /textitMedical Decision Making, 30, 672–684. http://dx.doi.or/10.1177/0272989X10369000.

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

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide (2nd ed.)*. Mahwah, NJ: Erlbaum.

Newstead, S. E., Pollard, P., Evans, J. S. B., & Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition*, *45*(3), 257–284.

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2015). Improving risk understanding across ability levels: Encouraging active processing with dynamic icon arrays. *Journal of Experimental Psychology: Applied*, *21*, 178–194. http://dx.doi.org/10.1037/xap0000045.

Peterson, C. R., & Miller, A. J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology, 70*, 117–121. http://dx.doi.org/10.1037/h0022023.

Pohl, R. F., & Erdfelder, E. (2017). Hindsight bias. In R. F. Pohl, R. F. Pohl (Eds.), *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory* (pp. 424–445). New York, NY, US: Routledge/Taylor & Francis Group.

Savani, K., & King, D. (2015). Perceiving outcomes as determined by external forces: The role of event construal in attenuating the outcome bias. *Organizational Behavior and Human Decision Processes, 130,* 136–146. http://dx.doi.org/10.1016/j.obhdp.2015.05.002.

Sedlmeier, P. & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130*, 380–400.

Sezer, O., Zhang, T., Gino, F., & Bazerman, M. H. (2016). Overcoming the outcome bias: Making intentions matter. *Organizational Behavior and Human Decision Processes, 137,* 13–26. http://dx.doi.org/10.1016/j.obhdp.2016.07.001.

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

Trippas, D., Handley, S. J., & Verde, M. F. (2013). The SDT model of belief bias: Complexity, time, and cognitive ability mediate the effects of believability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1393–1402. http://dx.doi.org/10.1037/a0032398.

Talboy, A. N., & Schneider, S. L. (2017). Improving accuracy on Bayesian inference problems using a brief tutorial. *Journal of Behavioral Decision Making, 30*, 373–388. http://dx.doi.org/10.1002/bdm.1949.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and* biases (pp. 153–160). New York, NY: Cambridge University Press.

Wagenmakers, E., Krypotos, A., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition, 40*, 145–160. http://dx.doi.org/10.3758/s13421-011-0158-0.