IMPROVING BAYESIAN REASONING: WHAT WORKS AND WHY?

EDITED BY: Gorka Navarrete and David R. Mandel
PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-745-3
DOI 10.3389/978-2-88919-745-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

IMPROVING BAYESIAN REASONING: WHAT WORKS AND WHY?

Topic Editors:
Gorka Navarrete, Universidad Diego Portales, Chile
David R. Mandel, York University, Canada

Cover portrait by Symen Veenstra (available at:http://enkeling.nl/2013/04/07/thomas-bayes-portrait/)

We confess that the first part of our title is somewhat of a misnomer. Bayesian reasoning is a normative approach to probabilistic belief revision and, as such, it is in need of no improvement. Rather, it is the typical individual whose reasoning and judgments often fall short of the Bayesian ideal who is the focus of improvement. What have we learnt from over a half-century of research and theory on this topic that could explain why people are often non-Bayesian? Can Bayesian reasoning be facilitated, and if so why? These are the questions that motivate this Frontiers in Psychology Research Topic.

Bayes' theorem, named after English statistician, philosopher, and Presbyterian minister, Thomas Bayes, offers a method for updating one's prior probability of an hypothesis H on the basis of new data D such that $P(H|D) = P(D|H)P(H)/P(D)$. The first wave of psychological research, pioneered by Ward Edwards, revealed that people were overly conservative in updating their posterior probabilities (i.e., $P(D|H)$). A second wave, spearheaded by Daniel Kahneman and Amos Tversky, showed that people often ignored prior probabilities or base rates, where the

priors had a frequentist interpretation, and hence were not Bayesians at all. In the 1990s, a third wave of research spurred by Leda Cosmides and John Tooby and by Gerd Gigerenzer and Ulrich Hoffrage showed that people can reason more like a Bayesian if only the information provided takes the form of (non-relativized) natural frequencies.

Although Kahneman and Tversky had already noted the advantages of frequency representations, it was the third wave scholars who pushed the prescriptive agenda, arguing that there are feasible and effective methods for improving belief revision. Most scholars now agree that natural frequency representations do facilitate Bayesian reasoning. However, they do not agree on why this is so. The original third wave scholars favor an evolutionary account that posits human brain adaptation to natural frequency processing. But almost as soon as this view was proposed, other scholars challenged it, arguing that such evolutionary assumptions were not needed. The dominant opposing view has been that the benefit of natural frequencies is mainly due to the fact that such representations make the nested set relations perfectly transparent. Thus, people can more easily see what information they need to focus on and how to simply combine it.

This Research Topic aims to take stock of where we are at present. Are we in a proto-fourth wave? If so, does it offer a synthesis of recent theoretical disagreements? The second part of the title orients the reader to the two main subtopics: what works and why? In terms of the first subtopic, we seek contributions that advance understanding of how to improve people's abilities to revise their beliefs and to integrate probabilistic information effectively. The second subtopic centers on explaining why methods that improve non-Bayesian reasoning work as well as they do. In addressing that issue, we welcome both critical analyses of existing theories as well as fresh perspectives. For both subtopics, we welcome the full range of manuscript types.

Citation: Navarrete, G., Mandel, D. R., eds. (2016). Improving Bayesian Reasoning: What Works and Why? Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-745-3

Table of Contents

frontiers in Psychology

EDITORIAL
published: 02 December 2015
doi: 10.3389/fpsyg.2015.01872

Editorial: Improving Bayesian Reasoning: What Works and Why?

David R. Mandel[1]* *and Gorka Navarrete*[2]*

[1] Department of Psychology, York University, Toronto, ON, Canada, [2] Laboratory of Cognitive and Social Neuroscience, Psychology Department, UDP-INECO Foundation Core on Neuroscience, Universidad Diego Portales, Santiago, Chile

Keywords: Bayesian reasoning, belief revision, risk communication, subjective probability, human judgment, individual differences, probabilistic judgment

This edited collection was motivated by an interest in understanding how to improve Bayesian reasoning. In that sense, the book before you is pragmatically and prescriptively oriented. Several of the papers address that challenge and some pick up on the important question of why certain factors work as well as they do. However, *Improving Bayesian Reasoning: What Works and Why* offers more than its editors had bargained for or its title suggests. Many papers offer methodological and conceptual insights that should help readers understand the *psychology* of Bayesian reasoning as practiced in cognitive science.

The book is comprised of 23 papers by 48 authors. The contributions are ordered by type: 10 original research articles first, followed by three reviews and 10 shorter essays. Foregoing an attempt to summarize each contribution in sufficient detail, let us simply draw out some observations about the collection.

ORIGINAL RESEARCH ARTICLES

This collection extends the base of original research on Bayesian reasoning in many important ways. Several papers offer further empirical evidence of the advantage of using visualized natural frequencies to communicate statistical information. Hoffrage et al. (2015b) show that the benefits of natural frequency representations in Bayesian tasks generalize from single- to multiple-cue cases and also to cases involving more than two hypotheses. Mandel (2015) shows that brief instruction in Bayesian reasoning using natural-frequency trees improves the coherence of intelligence analysts' posterior probability estimates. Binder et al. (2015) find that performance is improved when statistical information is communicated as natural frequencies instead of probabilities, and the natural-frequency format strengthens the facilitative effect of nested-set visualizations (i.e., tree diagrams and contingency tables) on Bayesian reasoning.

Other contributions identify where facilitative factors have their greatest impact. For instance, Hoffrage et al. (2015a) find that inexperienced business majors benefit more from natural-frequency formats than experienced business managers. Garcia-Retamero et al. (2015) address questions of *where* and *why* by showing that grid representations of natural frequencies facilitate Bayesian reasoning more strongly in medical patients with low numeracy, and that representational effects on reasoning are mediated by metacognitive judgment calibration. Hafenbrädl and Hoffrage (2015) go even further by parameterizing Bayesian skill using quantitative and qualitative factors that were free to vary across earlier studies. Finally, by triangulating choice and process data using an ecological sampling approach, Domurat et al. (2015) observe that many ostensibly Bayesian responses follow from use of an alternative statistical integration strategy.

The study of deduction had long been associated with reasoning from certain premises to certain conclusions. Yet Evans et al. (2015) and Cruz et al. (2015) venture into relatively new territory by

OPEN ACCESS

Edited and reviewed by:
Roberta Sellaro,
Leiden University, Netherlands

***Correspondence:**
David R. Mandel
drmandel66@gmail.com;
Gorka Navarrete
gorkang@gmail.com

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *12 November 2015*
Accepted: *19 November 2015*
Published: *02 December 2015*

Citation:
Mandel DR and Navarrete G (2015) Editorial: Improving Bayesian Reasoning: What Works and Why? Front. Psychol. 6:1872. doi: 10.3389/fpsyg.2015.01872

examining the quality of reasoners' *uncertain* deductions using coherence-based Bayesian metrics such as probabilistic validity. These papers capture the fundamental insight that, even in deduction, most arguments consist of uncertain premises from which uncertain conclusions are drawn.

Finally, the contribution by Douven and Schupbach (2015) is pragmatic in two unique senses. First, it causes us to reconsider whether Bayesianism is the most appropriate normative framework in some contexts. Second, in the tradition of the great American pragmatist Charles Sanders Peirce, it situates abduction within the normative fold. The authors argue that explanationist alternatives to Bayesianism not only withstand normative critiques, they also fare better descriptively.

REVIEW ARTICLES AND ESSAYS

The articles in this category draw out several dominant themes. First, the debate over natural frequencies vs. nested sets is passé. Although disagreement over the merits of the evolutionary account within which the original natural-frequency arguments were put forth linger, there is wide consensus that natural-frequency formats improve Bayesian performance by clarifying nested-set relations, which confers both representational and computational benefits (Brase and Hill, 2015).

Second, there has been a move away from the dual-systems account that emphasized System 1 sources of Bayesian error (Barbey and Sloman, 2007) toward a view that regards such errors as primarily due to representational and computational breakdowns in a problem-solving process, which occur even when explicit "System 2" processes are utilized (Johnson and Tubau, 2015; Sirota et al., 2015). For example, Juslin (2015) illustrates that Bayesian performance improves when computational requirements are shifted from multiplicative integration to additive integration. Likewise, Girotto and Pighin (2015) review studies showing that children and preliterate adults exhibit extensional reasoning that enables them to solve Bayesian problems provided they do not require explicit mathematical computation. The emerging view is further tempered by considerations of task characteristics, which are likely to alter the balance of implicit and explicit cognitive processes (Vallée-Tourangeau et al., 2015).

Whereas, most papers in this collection focus on Bayesian reasoners' performance, two refocus our attention on Bayesian communication by experts. Navarrete et al. (2014) make us consider how parents' decision-making about prenatal screening might be altered if they were given the positive predictive value (namely, the Bayesian value) of the initial screening test (which happens to be quite low) and also if parents received clear communications about the probabilistic risks of secondary invasive testing. Navarrete et al. (2015) generalize the argument, recommending that, where feasible, medical practitioners should give clients the relevant positive predictive values adjusted for their reference class. In short, clients should be relieved of computational burdens as far as possible so that they can focus on value-based decisions among available options.

Finally, several papers in this collection take the literature to task. Mandel (2014a) and McNair (2015) note that the definition of Bayesian reasoning in most psychological studies is mainly about information-integration performance. Few studies even require subjects to revise or update their beliefs! Others point to a lack of due attention to individual differences in reasoning and to the cognitive processes that lead to final estimates (Johnson and Tubau, 2015; McNair, 2015; Vallée-Tourangeau et al., 2015). Baratgin (2015) and Mandel (2014a) both take Bayesian researchers to task over their disregard of the subjectivist (and coherence-centered) foundations of Bayesianism.

However, attention to problems that have a temporal component is not lacking in this collection: Tubau et al. (2015) provide an insightful and comprehensive review of the Monty Hall Problem and Baratgin (2015) uses the two-player version of that problem to expose logical and terminological breakdowns in earlier theoretical analyses. Mandel (2014b) explores the perhaps even more complex Sleeping Beauty problem, which involves belief revision under conditions of asynchrony, to highlight how visual representations using quasi-logic trees can help clarify points of philosophical disagreement in the literature.

We hope readers will find this book informative, thought provoking, and of practical value.

AUTHOR CONTRIBUTION

Both authors wrote the manuscript.

ACKNOWLEDGMENTS

This work was supported by a grant from Comisión Nacional de Investigación Científica y Tecnológica (CONICYT/FONDECYT Regular 1150824 to GN).

REFERENCES

Baratgin, J. (2015). Rationality, the Bayesian standpoint, and the Monty-Hall problem. *Front. Psychol.* 6:1168. doi: 10.3389/fpsyg.2015.01168

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–297. doi: 10.1017/S0140525X07001653

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

Brase, G. L., and Hill, W. T. (2015). Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Front. Psychol.* 6:340. doi: 10.3389/fpsyg.2015.00340

Cruz, N., Baratgin, J., Oaksford, M., and Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Front. Psychol.* 6:192. doi: 10.3389/fpsyg.2015.00192

Domurat, A., Kowalczuk, O., Idzikowska, K., Borzymowska, Z., and Nowak-Przygodzka, M. (2015). Bayesian probability estimates are not necessary to make choices satisfying Bayes' rule in elementary situations. *Front. Psychol.* 6:1194. doi: 10.3389/fpsyg.2015.01194

Douven, I., and Schupbach, J. N. (2015). Probabilistic alternatives to Bayesianism: the case of explanationism. *Front. Psychol.* 6:459. doi: 10.3389/fpsyg.2015.00459

Evans, J. S. B. T., Thompson, V. A., and Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Front. Psychol.* 6:398. doi: 10.3389/fpsyg.2015.00398

Garcia-Retamero, R., Cokely, E. T., and Hoffrage, U. (2015). Visual aids improve diagnostic inferences and metacognitive judgment calibration. *Front. Psychol.* 6:932. doi: 10.3389/fpsyg.2015.00932

Girotto, V., and Pighin, S. (2015). Basic understanding of posterior probability. *Front. Psychol.* 6:680. doi: 10.3389/fpsyg.2015.00680

Hafenbrädl, S., and Hoffrage, U. (2015). Toward an ecological analysis of Bayesian inferences: how task characteristics influence responses. *Front. Psychol.* 6:939. doi: 10.3389/fpsyg.2015.00939

Hoffrage, U., Hafenbrädl, S., and Bouquet, C. (2015a). Natural frequencies facilitate diagnostic inferences of managers. *Front. Psychol.* 6:642. doi: 10.3389/fpsyg.2015.00642

Hoffrage, U., Krauss, S., Martignon, L., and Gigerenzer, G. (2015b). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Front. Psychol.* 6:1473. doi: 10.3389/fpsyg.2015.01473

Johnson, E. D., and Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Front. Psychol.* 6:938. doi: 10.3389/fpsyg.2015.00938

Juslin, P. (2015). Controlled information integration and Bayesian inference. *Front. Psychol.* 6:70. doi: 10.3389/fpsyg.2015.00070

Mandel, D. R. (2014a). The psychology of Bayesian reasoning. *Front. Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

Mandel, D. R. (2014b). Visual representation of rational belief revision: another look at the Sleeping Beauty problem. *Front. Psychol.* 5:1232. doi: 10.3389/fpsyg.2014.01232

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Front. Psychol.* 6:387. doi: 10.3389/fpsyg.2015.00387

McNair, S. J. (2015). Beyond the status-quo: research on Bayesian reasoning must develop in both theory and method. *Front. Psychol.* 6:97. doi: 10.3389/fpsyg.2015.00097

Navarrete, G., Correia, R., and Froimovitch, D. (2014). Communicating risk in prenatal screening: the consequences of Bayesian misapprehension. *Front. Psychol.* 5:1272. doi: 10.3389/fpsyg.2014.01272

Navarrete, G., Correia, R., Sirota, M., Juanchich, M., and Huepe, D. (2015). Doctor, what does my positive test mean? From Bayesian textbook tasks to personalized risk communication. *Front. Psychol.* 6:1327. doi: 10.3389/fpsyg.2015.01327

Sirota, M., Vallée-Tourangeau, G., Vallée-Tourangeau, F., and Juanchich, M. (2015). On Bayesian problem-solving: helping Bayesians solve simple Bayesian word problems. *Front. Psychol.* 6:1141. doi: 10.3389/fpsyg.2015.01141

Tubau, E., Aguilar-Lleyda, D., and Johnson, E. D. (2015). Reasoning and choice in the Monty Hall Dilemma (MHD): implications for improving Bayesian reasoning. *Front. Psychol.* 6:353. doi: 10.3389/fpsyg.2015.00353

Vallée-Tourangeau, G., Sirota, M., Juanchich, M., and Vallée-Tourangeau, F. (2015). Beyond getting the numbers right: what does it mean to be a "successful" Bayesian reasoner? *Front. Psychol.* 6:712. doi: 10.3389/fpsyg.2015.00712

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 14 October 2015
doi: 10.3389/fpsyg.2015.01473

Natural frequencies improve Bayesian reasoning in simple and complex inference tasks

Ulrich Hoffrage[1], Stefan Krauss[2], Laura Martignon[3] and Gerd Gigerenzer[4]*

[1] Faculty of Business and Economics (HEC Lausanne), University of Lausanne, Lausanne, Switzerland, [2] Mathematics Education, Faculty of Mathematics, University of Regensburg, Regensburg, Germany, [3] Institute of Mathematics, Ludwigsburg University of Education, Ludwigsburg, Germany, [4] Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany

Representing statistical information in terms of natural frequencies rather than probabilities improves performance in Bayesian inference tasks. This beneficial effect of natural frequencies has been demonstrated in a variety of applied domains such as medicine, law, and education. Yet all the research and applications so far have been limited to situations where one dichotomous cue is used to infer which of two hypotheses is true. Real-life applications, however, often involve situations where cues (e.g., medical tests) have more than one value, where more than two hypotheses (e.g., diseases) are considered, or where more than one cue is available. In Study 1, we show that natural frequencies, compared to information stated in terms of probabilities, consistently increase the proportion of Bayesian inferences made by medical students in four conditions—three cue values, three hypotheses, two cues, or three cues—by an average of 37 percentage points. In Study 2, we show that teaching natural frequencies for simple tasks with one dichotomous cue and two hypotheses leads to a transfer of learning to complex tasks with three cue values and two cues, with a proportion of 40 and 81% correct inferences, respectively. Thus, natural frequencies facilitate Bayesian reasoning in a much broader class of situations than previously thought.

Keywords: Bayesian inference, representation of information, natural frequencies, task complexity, instruction, fast-and-frugal trees, visualization

OPEN ACCESS

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
Håkan Nilsson, Uppsala University, Sweden
Miroslav Sirota, Kingston University, UK
Simon John McNair, Leeds University Business School, UK

****Correspondence:***
Ulrich Hoffrage, Faculty of Business and Economics (HEC Lausanne), University of Lausanne, Dorigny Batiment Internef, CH-1015 Lausanne, Switzerland
ulrich.hoffrage@unil.ch

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *24 May 2015*
Accepted: *14 September 2015*
Published: *14 October 2015*

Citation:
Hoffrage U, Krauss S, Martignon L and Gigerenzer G (2015) Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. Front. Psychol. 6:1473. doi: 10.3389/fpsyg.2015.01473

Introduction

After a positive hemoccult screening test, which signals hidden blood in the stool, a patient asks his doctor: "What does a positive result mean? Do I definitely have colon cancer? If not, how likely is it?" When 24 experienced physicians, including heads of departments, were asked this, their answers to the third question ranged between 1 and 99% (Hoffrage and Gigerenzer, 1998). All these physicians had the same information: a prevalence of 0.3%, a sensitivity of 50%, and a false positive rate of 3%. Bayes' rule shows that the actual probability of colon cancer given a positive result is about 5%. As this and subsequent studies have documented, most physicians do not know how to estimate the probability of cancer given the prevalence, sensitivity, and false positive rate of a test (Gigerenzer, 2014). This difficulty has also been observed in laypeople and attributed to some internal mental flaw, such as a general base rate neglect, the representative heuristic, or a general inability to reason the Bayesian way (e.g., Kahneman, 2011). Yet the experimental evidence has made it clear that the problem is not simply in our minds, but in the way the information

is presented. When Hoffrage and Gigerenzer (1998) gave another group of 24 physicians the same information in *natural frequencies* (see below), 16 of these could find the Bayesian answer, namely that a patient actually has cancer in only 1 out of 20 positive screening results. When given *conditional probabilities*, that is, the sensitivity and false alarm rate, only 1 out of 24 physicians could find the Bayesian answer, or anything close to it.

The positive effect of natural frequencies on Bayesian reasoning was first documented by Gigerenzer and Hoffrage (1995, 1999) and has since been confirmed in both numerous laboratory studies (e.g., Cosmides and Tooby, 1996; Brase, 2002, 2008) and applied research, including screening for Down syndrome (Bramwell et al., 2006), the interpretation of DNA evidence in court (Lindsey et al., 2003), and teaching children to reason the Bayesian way (Zhu and Gigerenzer, 2006). Thus, the earlier claim that people's cognitive limitations make them poor Bayesians (e.g., Kahneman and Tversky, 1972, repeated in Kahneman, 2011, and Thaler and Sunstein, 2008) is now known to be incorrect; it holds only when information is presented in probabilities. When presented in natural frequencies, by contrast, Bayesian performance increases substantially.

Yet there is a limitation to virtually all of these studies. Whether using conditional probabilities or natural frequencies, the experimental studies that have been conducted so far incorporated solely the simplest version of a Bayesian task—henceforth referred to as the *basic task*—which involves two hypotheses (such as colon cancer or no colon cancer) and a single cue (such as the hemoccult test) with two cue values (a positive or negative result). In 1998, Massaro questioned whether the facilitating effect of natural frequencies extends to more complex tasks that involve two or more cues. He conjectured that even in the case of two cues, "a frequency algorithm will not work" (p. 178). Although he did not test this claim, if true, it would severely limit the range of applications of natural frequencies. In this article, we experimentally test Massaro's claim, as well as whether the effect of natural frequencies generalizes to tasks involving three cues, three cue values, and three hypotheses.

This article has two parts. In the first, we outline the two paradigms for studying Bayesian reasoning, which use two different methodologies and have arrived at apparently contradicting conclusions concerning people's ability to reason the Bayesian way. One is a learning paradigm where probabilities are learned by sequentially observing events; the other is the classical textbook paradigm where people are assigned problems with specified conditional probabilities. We show that natural frequency representations are a kind of missing link between the two paradigms. In the second part, we report two studies. The first study tests whether the beneficial effect of natural frequencies generalizes to more complex Bayesian inferences, that is, to tasks containing more than two hypotheses, more than one cue, or cues with more than two values. The second study tests whether a short instruction in natural frequencies for a basic task (involving one dichotomous cue and two hypotheses) facilitates applying Bayesian reasoning to complex tasks. In the discussion we relate the present work to the fast-and-frugal heuristics program and to other interventions to boost performance in Bayesian inference tasks.

Paradigms to Study Bayesian Inferences: Probability Learning and Textbook Tasks

A Bayesian inference task is a task in which the probability $p(H|D)$ of some hypothesis H (e.g., cancer) given data D (e.g., a test result) has to be estimated. Two types of Bayesian inference tasks can be distinguished (Gigerenzer, 2015; Mandel, 2015; Sirota et al., 2015b): probability learning and textbook tasks.

Let us first consider probability learning tasks. Organisms learn the consequences of various behavioral responses in a probabilistic environment with multiple cues. Note that such a task ultimately requires behavioral responses in a specific situation. For instance, what should a bird do when it sees a movement in the grass? This situation can be conceived as a Bayesian inference task in which the behavioral response is based on a comparison of the probability that the movement of the grass (data, D) is caused by something that is dangerous (hypothesis, H) or by something that is not dangerous ($-H$). In the laboratory, a probability learning task involves the sequential encounter of pairs of events. In the case of two hypotheses (H and its complement $-H$) and two possible states of the world (data D observed or not), there are four possible pairs: $H\&D$, $H\&-D$, $-H\&D$, $-H\&-D$. To answer the Bayesian question "what is $p(H|D)$?" one needs to compare the two possibilities $D\&H$ and $D\&-H$ with respect to their probabilities. How likely is "grass movement due to dangerous cause (e.g., cat)" compared to "grass movement for some other non-dangerous reason (e.g., wind)"? How likely is "hemoccult test positive and patient has colon cancer" compared to "test positive for some other reason"? Transforming the odds of the two possibilities—one probability compared to the other—into a ratio amounts to dividing the first probability by the sum of both:

$$p(H|D) = \frac{p(D\&H)}{p(D)} = \frac{p(D\&H)}{p(D\&H) + p(D\&-H)} \quad (1)$$

where $p(H|D)$ stands for the posterior probability that the hypothesis H is true given the observed data D. Equation (1) is one form of Bayes' rule.

The probabilities relevant for Bayesian inferences can be learned via three paths: phylogenetic learning (natural selection of inherited instincts, i.e., evolutionary preparedness; Harlow, 1958), ontogenetic learning (e.g., classical and instrumental conditioning; Pearce, 1997), and, for some species, social learning (Richerson and Boyd, 2008). A major conclusion of the probability learning paradigm is that humans and animals are approximate Bayesians (Anderson, 1990; Gallistel, 1990; Chater et al., 2006; Chater and Oaksford, 2008).

Let us now turn to the second type of Bayesian inference tasks, textbook tasks. In their evolutionary history, humans have developed skills that other species have in some rudimentary form, but which humans master at a far superior level: social learning, instruction, and reasoning (Richerson and Boyd, 2008). These skills enable culture, civilization, science, and textbooks.

Moreover, they facilitate communication of probabilities, one of the many examples of how ontogenetic learning of probabilities can be supported by social learning (McElreath et al., 2013). Last but not least, they allow for the development of probability theory, which, in turn, offers a formal framework for evaluating hypotheses in light of empirical evidence. Even though the question of how this should be done is an ancient one, only since the Enlightenment have hypotheses been evaluated in terms of mathematical probability (Daston, 1988). Specifically, when evaluating an uncertain claim (i.e., hypothesis), the posterior probability of the claim can be estimated after new data have been obtained. One rigorous method for doing so was established by Thomas Bayes and, later, Pierre Simon de Laplace. The mathematical expression for updating hypotheses in light of new data is given in Equation (2):

$$p(H|D) = \frac{p(H)\,p(D|H)}{p(H)\,p(D|H) + p(-H)p(D|-H)} \quad (2)$$

where $p(H)$ and $p(-H)$ stand for the prior probabilities that the hypothesis (H) and its complement ($-H$), are true, and where $p(D|H)$ and $p(D|-H)$ stand for the likelihood of observing the data under these two different conditions. In signal detection theory, these two likelihoods are referred to as hit rate and false-alarm rate. In medical terms, the hit rate is the sensitivity of a diagnostic test and the false-alarm rate is the complement of the specificity of the test. Equation (2) formalizes how prior probabilities and likelihoods should be combined to compute the Bayesian posterior probability. Note that this equation is a variant of Equation (1) in which the two conjunctions, $p(D\&H)$ and $p(D\&-H)$, are broken into components. Strictly speaking, Equation (1), albeit a form of Bayes' rule, is not an equation that captures the updating of probabilities. Unlike Equation (2), Equation (1) does not describe the relationship between $p(H)$ and $p(H|D)$, simply because it does not include the term $p(H)$.

Social learning, probability theory, and Bayes' rule in the form of Equation (2) offer a new opportunity: to study Bayesian reasoning using textbook tasks with specified probabilities that do not need to be learned from experience. In contrast to the probability learning paradigm with its sequential input of observations, the textbook paradigm provides the information as a final tally (usually in numerical form). Whereas the most important cognitive ability required to solve a Bayesian task in the probability learning paradigm is frequency encoding (and memory), the most useful cognitive abilities in the textbook task paradigm are reasoning and calculation (for a discussion of Bayesian reasoning in textbox tasks adopting a problem-solving approach, see Johnson and Tubau, 2015). Note that the distinction between (Bayesian) behavior in the context of the probability learning paradigm and (Bayesian) reasoning in the context of the textbook paradigm is akin to Hertwig et al.'s (2004) distinction between decisions-from-experience and decisions-from-descriptions. But there are two kinds of descriptions within the textbook task paradigm: The statistical information can be presented in terms of either conditional probabilities or natural frequencies, which, as the introductory example illustrated, has quite opposite effects on reasoning.

Performance in Bayesian Textbook Problem Solving Depends on the Representation Format

It is striking to see the differences obtained by the two research paradigms (Gigerenzer, 2015; Mandel, 2015; Sirota et al., 2015b). Whereas the probability learning paradigm depicts humans and animals as approximate Bayesians (at least in the simple tasks studied), early research using the textbook paradigm arrived at a different conclusion. This discrepancy went mostly unnoticed because cross-references between the researchers in both paradigms have been rare. In their introductory note to the present special issue, Navarrete and Mandel (2015) distinguish three waves in the history of this research using the textbook paradigm. The first wave was marked by Edwards (1968) with his urns-and-balls problems. In the vignettes of these problems, prior probabilities [i.e., $p(H)$ and $p(-H)$] were communicated but no likelihoods [i.e., $p(D|H)$ and $p(D|-H)$]—although the sample information that was given instead (e.g., 4 blue balls and 1 red ball) potentially allowed for calculating the corresponding likelihoods. Edwards (1968) found that if people have to update their opinions, they change their view in the direction proposed by Bayes' rule. However, he also reported that people are "conservative Bayesians" in the sense that they do not update their prior beliefs as strongly as required by Bayes' rule.

A study by Eddy (1982) illustrates the second wave of research. The question he asked was: Do experts reason the Bayesian way? Eddy found that physicians' judgments did not follow Bayes' rule when solving the following type of task (a prototypical Bayesian situation):

> *The probability of breast cancer is 1% for a woman at age 40 who participates in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?*

According to Bayes's rule, the answer is 7.8%, which can be obtained by inserting the given information into Equation (2). Yet Eddy (1982) reported that 95 out of 100 physicians estimated this probability to be between 70 and 80%. He argued that these physicians confused the conditional probability of breast cancer given a positive mammogram with that of a positive mammogram given breast cancer. To explain the failure of Bayesian reasoning, Kahneman and Tversky (1972) suggested the "representativeness heuristic," although it remains unclear whether the heuristic concurs with Eddy's explanation because this "one-word explanation" (Gigerenzer, 1996, p. 594) has never been defined and formalized (see Gigerenzer and Murray, 1987). Be that as it may, Kahneman and Tversky (1972) concluded: "In his evaluation of evidence man is apparently not a conservative Bayesian: he is not Bayesian at all" (p. 450).

Whereas the second wave attributed failure in Bayesian reasoning to flawed mental processes, a third wave starting in the mid-1990s (Gigerenzer and Hoffrage, 1995, 1999; Cosmides

and Tooby, 1996) showed experimentally that much of the problem lies in how risk is represented. Specifically, Gigerenzer and Hoffrage established that it is not Bayesian reasoning *per se* that is difficult but rather the format of information provided to the participants. In Eddy's (1982) task, quantitative information was provided in conditional probabilities. Gigerenzer and Hoffrage (1995) showed that such a representation format makes the computation of the Bayesian posterior probability more complicated than with natural frequencies. Natural frequencies result from natural sampling and have historically been the "natural" input format for the human mind (Kleiter, 1994; Gigerenzer and Hoffrage, 1999, pp. 425–426). Presenting the information in Eddy's mammography task in terms of natural frequencies yields the following description:

> *10 out of every 1000 women at age 40 who participate in routine screening have breast cancer. 8 out of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at age 40 who got a positive mammography in a routine screening. How many of these women do you expect to actually have breast cancer?*

Answering this question amounts to solving Equation (3):

$$p(H|D) = \frac{f(D\&H)}{f(D)} = \frac{f(D\&H)}{f(D\&H) + f(D\&\text{–}H)} \qquad (3)$$

where *f(D&H)* stands for the natural frequency of joint occurrences of *D* and *H*, *f(D&–H)* stands for the natural frequency of joint occurrences of *D* and *–H*, and *f(D)* for their sum. In the mammography problem, these two joint occurrences are 8 and 95 (out of 1000 women), respectively, and hence there are, in sum, 103 women who get a positive mammogram. Of 103 women who get a positive mammogram, 8 actually have breast cancer. This relative frequency of 8/103 corresponds to a posterior probability of 7.8%, the number that we already computed using Equation (2). Note that natural frequencies result from drawing N objects (e.g., 1000 in the above example) at random from a larger population (or from taking the entire population). Any decomposition of this sample of size N contains natural frequencies, which can be interpreted only in relation to each other and in relation to the total sample size N. Attempts to illustrate what natural frequencies are by simply naming "1 of 10" as an example and in isolation from any other number of a natural frequency tree misses this important point.

When information has been presented in terms of natural frequencies, almost half of Gigerenzer and Hoffrage's (1995) student participants found the Bayesian answer. Among 160 gynecologists, the proportion of Bayesian answers increased from 21 to 87% for probabilities and natural frequencies, respectively (Gigerenzer et al., 2007). The beneficial effect of natural frequency representations has been replicated with experienced physicians (Hoffrage and Gigerenzer, 1998; Bramwell et al., 2006), patients (Garcia-Retamero and Hoffrage, 2013), judges (Hoffrage et al., 2000), and managers (Hoffrage et al., 2015), and has been used to design tutorials on Bayesian reasoning (Sedlmeier and Gigerenzer, 2001; Kurzenhäuser and Hoffrage, 2002).

Textbook problems with information provided in terms of natural frequencies are in fact close to the probability learning paradigm [see the similarity between Equations (1) and (3)]. In contrast, textbook problems with information provided in terms of probabilities do not bear much resemblance to this paradigm [note the difference between Equation (2), with its three pieces of information, and Equation (1), with its two pieces of information]. Natural frequencies are related to the probability learning paradigm because they are the final tally that result from what has been called "natural sampling" (Kleiter, 1994) which, in turn, can be conceived as the process of sequentially observing one event after the other in a natural environment. In other words, natural sampling is the process underlying experiential learning—the paradigm in which humans and animals tend to perform well (Hasher and Zacks, 1979; Gallistel, 1990). Thus, it is no surprise that the beneficial effect of natural frequency representations could be found even for 4th and 5th graders (Zhu and Gigerenzer, 2006; Gigerenzer, 2014; Multmeier, unpublished manuscript; see also Till, 2013). The comparison between Equations (2) and (3) shows why natural frequencies facilitate Bayesian inference. It simplifies computation of the posterior probability: The representation does part of the computation (Gigerenzer and Hoffrage, 2007; Hill and Brase, 2012; Brase and Hill, 2015).

In subsequent work, the power of representation formats has been discussed in a wider context that also embraces important issues such as trust, transparency, or institutional design, to name a few (see Gigerenzer, 2002, 2014; Gigerenzer et al., 2007; Gigerenzer and Gray, 2011). As a consequence of all this research, of various activities to propagate it, and of the desire and pressure to improve Bayesian inference in several domains, the use of natural frequencies is recommended by major evidence-based medical societies, including the Cochrane Collaboration (Rosenbaum et al., 2010), the International Patient Decision Aid Standards Collaboration (Trevena et al., 2012), the Medicine and Healthcare Products Regulatory Agency (the United Kingdom's equivalent to the Food and Drug Administration; see Woloshin and Schwartz, 2011), and the Royal College of Obstetricians and Gynecologists (2008). Moreover, natural frequencies are used in some of the most important school textbooks and in textbooks for future teachers of stochastics in school in the German speaking countries (Martignon, 2011), and they are already part of the school syllabus in the United Kingdom (Spiegelhalter and Gage, 2014).

Yet, as mentioned in the introduction, these developments are severely limited by the fact that up to now, the studies on which they are based used only simple versions of Bayesian tasks with one dichotomous cue and two hypotheses.

Types of Bayesian Inference Tasks: The Basic Task and Complex Tasks

There is one important difference between real-life probability learning tasks and textbook problem solving. Compared to most

real-life situations, the textbook problems in the literature on Bayesian reasoning are relatively simple. The vast majority of them involve two hypotheses and one dichotomous cue. As mentioned before, we refer to such a task as a basic task. Many real-life situations, in contrast, are more complex. We see three ways in which the basic task can be extended; **Figure 1** depicts the basic task (**Figure 1A**) and these extensions (**Figures 1B–E**).

One extension involves a situation with a cue having more than two levels (**Figure 1B**). In fact, many variables are polychotomous. Others may even be continuous and have been divided, for various reasons, into several categories by using cutoffs. For instance, mammograms obtained in a screening program are not simply positive or negative but depict breast cancers that vary in size, shape, or density, the fact of which led to the BI-RADS classification that distinguishes multiple categories. Generally speaking, for polychotomous cues, there is not only one hit rate, $p(D|H)$, and one false-alarm rate, $p(D|{-}H)$, but there are, both for H and for $-H$, as many likelihoods as there are categories for the data: $p(D_1|H)$, $p(D_2|H)$,..., $p(D_n|H)$ and $p(D_1|{-}H)$, $p(D_2|{-}H)$,..., $p(D_n|{-}H)$, respectively. Correspondingly, there are as many posterior probabilities (with their complements) as there are data categories: $p(H|D_1)$, $p(H|D_2)$,..., $p(H|D_n)$. **Figure 1B**, illustrates a situation used in the studies reported below, namely a cue that has either a positive, a negative, or an unknown value.

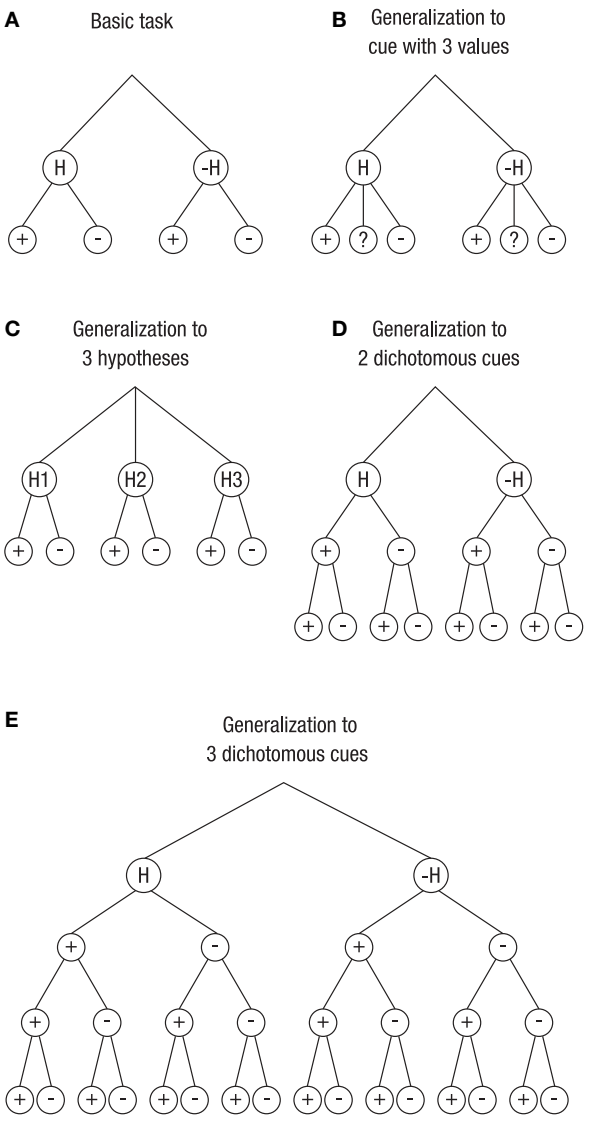

FIGURE 1 | Generalization of the basic Bayesian inference task (with two hypotheses and one dichotomous cue; A) to more complex tasks (B–E). The layers below the hypotheses depict the cue values (or data). Unknown cue values are denoted as "?" **(B)**. For a pair of two hypotheses (one being the complement of the other), these are denoted as H and –H **(A,B,D,E)**, and for a triple of hypotheses, they are denoted as H1, H2, and H3 **(C)**.

Figure 1C depicts a situation with three hypotheses. For instance, a fever may have many different causes, so no physician will prescribe a drug on the basis of fever alone but will ask further questions to assess the probabilities for multiple candidate reasons. Accordingly, while there are two likelihoods in the basic task—the hit rate, $p(D|H)$, and the false-alarm rate, $p(D|{-}H)$—there are now as many conditional probabilities for the complex task as there are hypotheses: $p(D|H_1)$, $p(D|H_2)$,..., $p(D|H_m)$. The same applies for the posterior probability, where there are no longer just two, $p(H|D)$ and $p({-}H|D)$, but rather as many probabilities as there are hypotheses, $p(H_1|D)$, $p(H_2|D)$,..., $p(H_m|D)$.

Finally, **Figures 1D,E** depict a situation with more than one cue. Asking for more information after the doctor has learned that the patient has fever amounts to inspecting more cues or performing additional tests.

How do natural frequencies affect Bayesian performance in these three complex tasks? Whereas Gigerenzer and Hoffrage (1995) left open whether the beneficial effect of natural frequencies can be generalized to more complex tasks, Massaro (1998) questioned, as mentioned before, this possibility for situations with more than one cue. Unlike in **Figure 1D**, he did not add one layer per cue but instead arranged the possible combinations of cue values—for a situation with two cues—in one single layer. That is, directly under the node depicting that "hypothesis H is true," he placed four branches depicting the four possible combinations of two dichotomous cues: $+C_1\&+C_2$, $+C_1\&-C_2$, $-C_1\&+C_2$, and $-C_1\&-C_2$ (where + and – denote positive and negative cue values for the two cues C_1 and C_2). Moreover, he argued that "it might not be reasonable to assume that people can maintain exemplars of all possible symptom configurations" (p. 178). However, he did not provide any empirical evidence for this claim. We fill this gap by analyzing how participants perform in complex Bayesian tasks dependent on whether information is provided in terms of probabilities or natural frequencies.

Study 1: Bayesian Inferences in Complex Tasks

Method

Participants were advanced medical students ($N = 64$) of the Free University of Berlin. Each of them was asked to

work on four medical diagnostic tasks. Task 1 was a Bayesian task corresponding to **Figure 1B**, in which we extended Eddy's mammography task by adding unclear test results. Task 2 was a Bayesian task corresponding to **Figure 1C**, where a test could detect two diseases, namely Hepatitis A and Hepatitis B. Tasks 3 and 4 were Bayesian tasks with two and three cues, corresponding to **Figure 1D** and **Figure 1E**, respectively. In Task 3, breast cancer had to be diagnosed based on a mammogram and an ultrasound test. In Task 4, an unnamed disease had to be diagnosed on the basis of three medical tests, simply named Test 1, Test 2, and Test 3. The participants could work on the four tasks at their own pace, which took them, on average, about 1 h in total.

Each participant received the statistical information for two of the four tasks in probabilities and the other two in natural frequencies. As an illustration, **Table 1** displays the two different versions (probability version vs. natural frequency version) of Task 3. The exact formulations of Tasks 1, 2, and 4 can be seen in Appendix I (Supplementary Material). Note that for Tasks 3 and 4, not all natural frequencies on the lowest layer (i.e., for all combinations of the two and three cues, respectively) were stated, but only those for which all tests were positive. Besides requesting a numerical answer to each of these four tasks, we also asked the participants to make notes and to justify their answers so that we could better understand their reasoning processes. Pocket calculators were not allowed. Following Gigerenzer and Hoffrage (1995), we classified a response as Bayesian if it was either the exact Bayesian solution or rounded to the next full percentage point.

Figure 2 illustrates the frequency tree for the information provided in Task 3. Note, however, that the participants in Study 1 were neither presented with trees nor told to construct them; rather, they had to solve the task based on the wording alone[1].

[1]Note that the wording of the probability version of Task 3 is mute on the question of whether or not the two tests are independent of each other. Even though each of the two tests is dependent on the disease, they are indeed independent of each other for any level of the variable *disease*, and we anticipated that our participants (advanced medical students) knew this. An analysis of the participants' protocols revealed that all of them implicitly made this assumption. Participants' intuitive assumption of independence was also found in Task 4, where information on three unnamed tests was provided. This finding is in accordance with the finding of Waldmann and Martignon (1998) that people assume conditional independence between cues as long as there is no explicit evidence suggesting dependency.

Results

Figure 3 displays the percentage of correct Bayesian inferences for each of the four tasks. In all of the tasks, replacing probabilities with natural frequencies helped the medical students make better inferences. The percentage of correct Bayesian inferences averaged across the probability versions of the four tasks was 7%; across the natural frequency versions it was 45%. Natural frequencies were most effective in Task 1, where the difference in terms of participants' performance between the natural frequency and the probability version was 59% – 1% = 58 percentage points. In the other three tasks, the increase in participants' performance from the probability versions to the natural frequency versions was about 30 percentage points. A comparison of Tasks 3 and 4 suggests that, for both the probability and the natural frequency versions, it did not matter whether information was provided on two or on three cues or whether this information referred to named or unnamed tests and diseases.

Discussion

Study 1 showed that natural frequencies facilitate Bayesian reasoning in four complex tasks, relative to probabilities. How does the effect of natural frequencies on solving complex tasks compare to their effect on solving a basic task? One might expect that Bayesian performance in complex tasks decreases in both formats and—due to bottom effects for the probability format—that the facilitating effect of natural frequencies is less pronounced for complex tasks. However, that does not seem to be the case. Both in the present Study 1 and in Gigerenzer and Hoffrage (1995, Study 1), who used the same kind of problems, albeit for basic tasks, the average increase in performance when given natural frequencies rather than probabilities was similar, 38 percentage points in the present study (averaged across the 4 tasks) and 30 percentage points in their study. Thus, the comparison between these studies suggests the surprising conclusion that increased complexity may not decrease the effect of natural frequencies much. Whether that also holds for levels of complexity that go beyond those studied here is unknown.

Study 2: Transfer Learning

Sedlmeier and Gigerenzer (2001) and Kurzenhäuser and Hoffrage (2002) have shown that the beneficial effect of

TABLE 1 | Study 1, Task 3: A generalization of the basic Bayesian task to a more complex task with two cues (corresponding to Figure 1D).

Probability version	Natural frequency version
The probability of breast cancer is 1% for a woman at age 40 who participates in routine screening. If a woman has breast cancer, the probability is 80% that she will have a positive mammogram. If a woman does not have breast cancer, the probability is 9.6% that she will also have a positive mammogram. If a woman has breast cancer, the probability is 95% that she will have a positive ultrasound test. If a woman does not have breast cancer, the probability is 4% that she will also have a positive ultrasound test.	100 out of every 10,000 women at age 40 who participate in routine screening have breast cancer. 80 out of every 100 women with breast cancer will receive a positive mammogram. 950 out of every 9900 women without breast cancer will also receive a positive mammogram. 76 out of 80 women who had a positive mammogram and have cancer also have a positive ultrasound test. 38 out of 950 women who had a positive mammogram, although they do not have cancer, also have a positive ultrasound test.
What is the probability that a woman at age 40 who participates in routine screening has breast cancer, given that she has a positive mammogram and a positive ultrasound test?	How many of the women who receive a positive mammogram and a positive ultrasound test do you expect to actually have breast cancer?

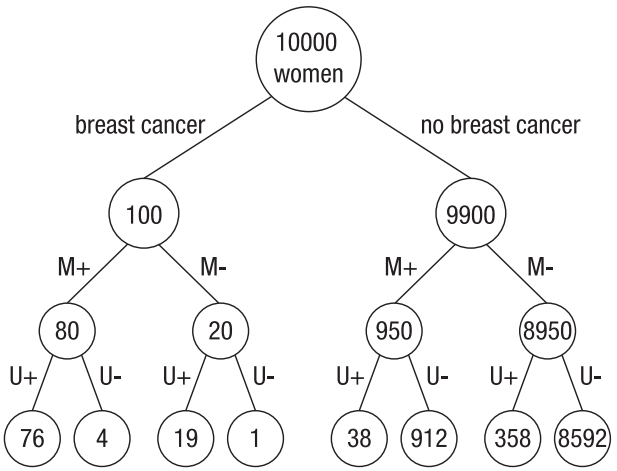

FIGURE 2 | Visual representation of the information provided in the natural frequency version of Task 3 of Study 1. "M" and "U" denote mammography and ultrasound Test, and "+" and "−" denote positive and negative test results, respectively.

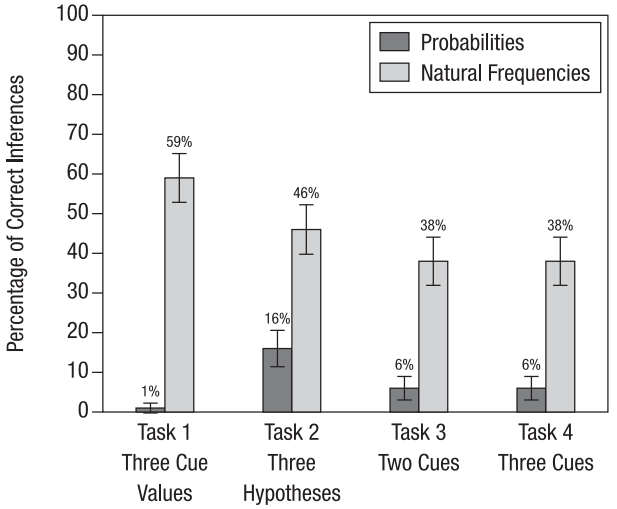

FIGURE 3 | Percentage of correct inferences in the four tasks used in Study 1. The bars display standard errors.

presenting information in natural frequencies can be enhanced by teaching people to use this representation. In one of their studies, Sedlmeier and Gigerenzer gave two groups of participants a computerized tutorial: One group was taught how to represent probabilities in terms of natural frequencies, supported by two visual aids—frequency grid and frequency tree (representation training); the other was taught Bayes' rule for probabilities (rule training). After training, participants in each group were tested on tasks in which the statistical information was always provided in terms of probabilities. The immediate learning success for the representation training group was an improvement from 10 to 90% Bayesian answers, compared to an improvement from 0% to about 65% for the rule training group. More important, the improvement in the representation training condition was stable over time. Even 5 weeks after training, the performance of the participants who had learned to use natural frequencies remained a high 90%, whereas the performance of the group with rule training dropped to about 20%. These results were obtained for basic Bayesian tasks.

In Study 2 we addressed the question of whether in place of a computerized training program, a simple written instruction on how to solve a basic task could improve participants' ability to solve complex tasks. Extending Study 1, which investigated whether the beneficial effect of presenting information in terms of natural frequencies could also be observed for complex Bayesian tasks, Study 2 investigated whether the beneficial effect of teaching Bayesian reasoning by training representations with a basic task can also be observed when participants are later tested with complex Bayesian tasks (for which they did not receive any training).

Method

We recruited advanced medical students ($N = 78$) from Berlin universities (none of them was a participant in Study 1). In the first step, each participant received a two-page instruction sheet on how to solve the mammography task, that is, a basic task with two hypotheses and one dichotomous cue. There were three different instructions, and participants were randomly assigned to one of them [all three instructions are shown in Appendix II (Supplementary Material)]. For Group 1, the mammography task was presented in terms of probabilities, and participants were shown how they could solve it by inserting the probabilities into Bayes' rule. For Group 2, the mammography task was presented in terms of probabilities, but here participants were instructed how to translate the probabilities into natural frequencies, how to place these frequencies into a tree, and how to determine the answer from this tree. For Group 3, the mammography task was presented in terms of natural frequencies (but no probabilities were provided), and these participants also received instructions on how to solve it by means of the frequency tree.

After studying their instruction sheet, participants were given two test tasks—the same that were used in Task 1 (one cue with three cue values) and Task 3 (two cues with two cue values each) in Study 1. Participants of Group 1 and 2 received probability versions of these tasks, and participants of Group 3 received the natural frequency version. The instruction sheet was at their disposal while working on the complex tasks. **Table 2** summarizes the design of Study 2.

Results

Figure 4 displays the percentages of Bayesian inferences in Tasks 1 and 3 separately for the three experimental groups. In both tasks, participants' performance was about the same, which suggests that the differences found in Study 1 disappear when there is an instruction on the basic task.

For the basic task, participants in Group 1 learned how to insert probabilities into Bayes' rule. Then they were tested on whether this training generalizes to applying Bayes' rule to more complex tasks in which information is presented in probabilities. Compared to Groups 2 and 3, this group performed worst when

TABLE 2 | Experimental design in Study 2: Three ways to instruct participants to solve the mammography task.

	Group 1 ($N = 27$)	Group 2 ($N = 25$)	Group 3 ($N = 26$)
Basic task used for instruction	Mammography task, formulated in terms of probabilities	Mammography task, formulated in terms of probabilities	Mammography task, formulated in terms of natural frequencies
Solution explained in instruction	How to insert probabilities into Bayes' rule	(a) How to translate probabilities into natural frequencies	How to place these natural frequencies into a frequency tree and to extract the correct answer
		(b) How to place these natural frequencies into a frequency tree and to extract the correct answer	
Complex tasks tested	Tasks 1 and 3 of Study 1 (both tasks in probabilities)	Tasks 1 and 3 of Study 1 (both tasks in probabilities)	Tasks 1 and 3 of Study 1 (both tasks in natural frequencies)

For details, see Appendix II in Supplementary Marerial.

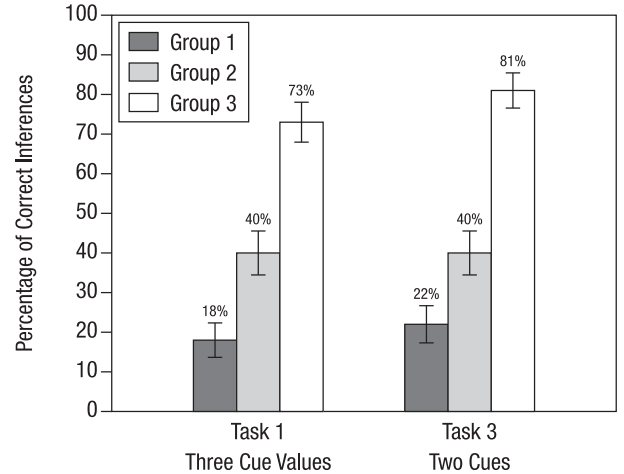

FIGURE 4 | Percentage of correct inferences for Tasks 1 and 3, depending on how participants were instructed to solve Bayesian inference tasks in Study 2. The bars display standard errors. Tasks 1 and 3 are the same test tasks as those that were used for Study 1. The three instruction conditions are summarized in **Table 2** and can be seen in full length in Appendix II (Supplementary Material).

confronted with complex Bayesian tasks (18% for Task 1 and 22% for Task 3). Nonetheless, their percentage of Bayesian inferences was substantially higher compared to that of participants of Study 1 for the same tasks (1% for Task 1 and 6% for Task 3; see **Figure 3**). Hence, we can conclude that the instruction had a positive effect: At least some of the participants managed to extend Bayes' rule to a more complex task involving an unclear test result (which amounts to adding a corresponding term to the denominator of Equation 2) and to a more complex task involving the results of two different tests (which amounts to applying Bayes' rule twice, that is, first computing the posterior probability after the first test result became known, and then using this probability as a prior probability to compute the posterior probability after the result of the second test became known).

Participants in Group 2 had learned, for the basic task, how to translate probabilities into natural frequencies. In spite of also being tested on tasks with information presented in terms of probabilities, 40% of participants in Group 2 obtained the correct solutions (this percentage happened to be identical for Tasks 1 and 3). These participants arrived at these solutions by performing the following steps: First, they correctly translated five probabilities (rather than three, as was the case for the basic task) into natural frequencies. To construct a corresponding tree they added nodes to the tree they had seen in the instruction. For Task 1 they had to add two nodes on the lowest layer (as can be seen when comparing **Figure 1A** and **Figure 1B**), and for Task 3 they had to add an additional layer for the outcomes of the ultrasound test (as can be seen when comparing **Figure 1A** and **Figure 1D**). From these modified trees they finally extracted the frequencies needed for the Bayesian solutions in the form of "Laplacian proportions," that is, the ratio of relevant cases divided by the total number of cases.

The participants of Group 3 were the only ones who were trained and tested with natural frequencies. This instruction method led to a high performance rate of 73% (Task 1) and 81% (Task 3). In contrast to Group 2, participants of Group 3 only needed to extend frequency trees; no translation of probabilities into frequencies was required. Recall that without prior instruction on the basic task, performance on the same two tasks was lower, 59 and 38%, respectively (Study 1). When comparing the performance gain for Task 1 (from 59% in Study 1, without instruction, to 73% in Study 2, with instruction) with the corresponding performance gain for Task 3 (a rise from 38 to 81%), it becomes obvious that instructions based on frequency representations affected the two types of generalizations differentially. Analyzing participants' protocols confirmed this pattern: Participants found it easier to take the tree from the basic task and to add another layer than to add nodes within a layer. In other words, generalizing the basic task (**Figure 1A**) to Task 3 (**Figure 1D**) seemed to be more intuitive for the participants than generalizing it to Task 1 (**Figure 1B**).

Discussion

Previous studies have established the usefulness of teaching how to represent probability information in terms of natural frequencies (Kurzenhäuser and Hoffrage, 2002; Sedlmeier and Gigerenzer, 2001; Ruscio, 2003; Sirota et al., 2015a). Study 2

extends these findings by showing that a simple instruction on how to solve a basic Bayesian task can amplify performance in complex tasks. The highest levels were obtained when both the trained task and the tested task were consistently formulated in terms of natural frequencies. That is, it is largely sufficient to instruct people in using natural frequencies in the basic task in order to ensure a generalization to and solution of complex tasks, as long as the information in both cases is in natural frequencies.

General Discussion

This paper has two results, one conceptual and one empirical. **Figure 1** shows how the natural frequency tree for the basic task (**Figure 1A**) can be generalized to various complex Bayesian tasks. As these trees (displayed in **Figures 1B–E**) demonstrate, the possibility of communicating statistical information in terms of natural frequencies is not restricted to the basic task with one dichotomous cue for inferring which of two hypotheses is true. Being able to generalize from these trees is important because in many real-life situations such as medical diagnosis or court trials, information is not dichotomous, several (rather than only one) pieces of evidence are available, and/or more than two hypotheses are considered.

With Study 1, we have empirically shown that, despite the trees for complex tasks having more branches than in the tree for the basic task, the facilitating effect of natural frequencies is essentially in the same order of magnitude as in previous studies using the basic task. Study 2 showed that instructing people how to use natural frequencies to solve the basic task was helpful for solving complex Bayesian tasks. Apparently, the best method is to instruct directly how to reason with natural frequencies and also to test people on natural frequencies. Instruction adds to the mere effect of representation demonstrated in Study 1. In contrast to claims made in the literature (Massaro, 1998), each of our studies show that the power of natural frequencies generalizes to complex tasks. In the remainder of this paper, we will discuss the power (and limits) of natural frequencies and that of instructions.

Power (and Limits) of Natural Frequencies in Complex Tasks

This study has shown that the natural frequency approach to Bayesian reasoning is powerful enough to be generalized to complex tasks and to allow for good performance despite increasing numbers of cues and cue values. How do natural frequencies support reasoning? Gigerenzer and Hoffrage (1995) demonstrated in detail that natural frequencies reduce the number of computational steps necessary for Bayesian inference and derived seven specific results, including that relative frequencies do not simplify the computation. Subsequent work has used different terms for the same explanation: the subset principle, set inclusion, or the nested-set hypothesis (for a discussion of these terms and their relationship to natural frequencies, see Hoffrage et al., 2002; Brase, 2007; Ayal and Beyth-Marom, 2014). Moreover, Ayal and Beyth-Marom also quantified the computational simplification and counted the mental steps or elementary information processes as a measure of the cognitive effort required to complete the task (for a similar analysis, see Johnson and Tubau, 2015).

Extending this analysis to complex tasks is straightforward and reveals that natural frequencies require less cognitive effort not only for basic tasks but also for complex tasks. However, even natural frequencies require computation and effort. Hence it does not come as a surprise (1) that for tasks using natural frequencies, the proportion of Bayesian inferences is less than 100% and (2) that variables related to participants' computational abilities can account for variance in Bayesian performance. For instance, performance in Bayesian inference tasks—both for probability and natural frequency representations—is correlated with numeracy (Chapman and Liu, 2009; Johnson and Tubau, 2015), numerical skills (Tubau, 2008), and fluid cognitive ability and thinking disposition (Sirota et al., 2014) (for a discussion of individual differences in Bayesian reasoning, see Brase and Hill, 2015).

At the same time, natural frequency representations have their limits. As mentioned earlier, Massaro (1998) argued that "a frequency algorithm will not work" because "it might not be reasonable to assume that people can maintain exemplars of all possible symptom configurations" (p. 178). We have meanwhile seen that in a textbook task, half (and with instruction, three quarters) of our participants were able to process the statistical properties of three cues in a Bayesian way when this information was represented in terms of natural frequencies. Notwithstanding this result, we share Massaro's concern that at some point humans are no longer able to store the frequencies for all possible conjunctions of cues in memory. In fact, in a situation with 10 dichotomous cues, the corresponding frequency tree would carry 2048 natural frequencies on the lowest layer, and with 20 cues this number would be over 2 million. It may nonetheless be possible to learn the statistical relationships between hypotheses and cues as the number of cues grow larger—after all, for two hypotheses and a dichotomous cue there are only four proportions (or probabilities) that are relevant and need to be learned: $p(D|H)$, $p(D|\neg H)$, $p(H|D)$, $p(H|\neg D)$. Learning the statistical relationships for *conjunctions* of cues, however, is a huge challenge because the number of relevant proportions would no longer grow linearly with the number of cues (four per cue) but instead exponentially.

As the number of cues grows larger, the difference between real-life settings and textbook tasks becomes increasingly important. Whereas it is difficult, if not impossible, to memorize and manage the relevant information in a real-life setting, which corresponds to a probability learning paradigm, it is possible to represent the natural frequencies required for Bayesian inferences in a textbook task. But even natural frequency representations in textbook tasks have their limits. These may not yet be reached for three cues, as our empirical findings reported above suggest, but draw nearer as the tree grows larger. The two major limits are practical feasibility and robustness. First, practical feasibility is hampered by the sheer amount of information that needs to be communicated—recall that a frequency tree for two hypotheses and 10 (20) dichotomous cues would have 2048 (>2,000,000) natural frequencies on the lowest layer. Second, and relatedly, for many real-life applications the

number of observations for a particular combination of cues will most likely be relatively small. Because of the resulting estimation error, the Bayesian inferences may have fairly wide confidence intervals and may thus not be very robust.

Fast-and-frugal Trees

What tools remain for the boundedly rational human mind (and for animals) in complex situations with a vast number of cues? We assume that the human mind is equipped with an adaptive toolbox containing simple heuristics that allow "fast-and-frugal" decisions, even in highly complex environments (Gigerenzer et al., 1999, 2011; Gigerenzer and Selten, 2001; Todd et al., 2012; Hertwig et al., 2013). These simple heuristics are helpful when making inferences in situations under limited time, with limited knowledge, and within our cognitive and computational constraints. One of the characteristics of these simple heuristics is that they reduce information intake and processing. Complexity—and note that this is the direction in which we extended the basic task—can be reduced tremendously by assuming conditional independence between cues, which is exactly what participants seem to do unless they have strong evidence speaking against this assumption (Waldmann and Martignon, 1998; Martignon and Krauss, 2003). To the extent that this assumption is justified, it is no longer necessary to store the millions of possible conjunctions of 20 dichotomous cues in memory, but it would be sufficient to represent the predictive power of a cue independent of the other cues.

The reduction of complexity can be achieved in many ways. Radically pruning a natural frequency tree for many cues while maintaining all cue information converts it into a so called *fast-and-frugal tree*—which is one of the heuristics analyzed by the Center for Adaptive Behavior and Cognition at the Max-Planck Institute for Human Development in Berlin (Martignon et al., 2003). **Figure 5** shows an example of such a classification tree, based on Green and Mehr (1997), for classifying patients as at high or low risk for heart disease. In **Figure 5A**, the full natural frequency tree for three cues is exhibited. Note that this tree displays the hypotheses (high risk vs. low risk of heart attack) no longer at the second layer, as the trees in **Figure 1** do, but at the very lowest layer. Whereas the trees in **Figure 1** are the usual natural frequency trees that communicate data given a hypothesis, the tree in **Figure 5A** displays natural frequencies *after* Bayesian updating, which, in turn, enables the classification of patients based on symptoms. Note that the trees in **Figure 1** and **Figure 5** carry natural frequencies (for a direct comparison of these two forms of grouping a given set of natural frequencies, see Hoffrage et al., 2015, Figures 1B,C).

The tree in **Figure 5A** can be radically pruned. The resulting fast-and-frugal tree, exhibited in **Figure 5B**, is "fast and frugal" according to the definition given in Martignon et al. (2008): At each node of the tree, the choice is either to stop further information acquisition and make a diagnosis or to collect more information. Specifically, in a first step, all 89 patients are checked for elevated ST segment in their electrocardiogram. If the answer is positive (ST+), they ($n = 33$) are classified as high risk, without considering any further information. The remaining 56 patients are checked for chest pain as the main symptom. If the answer is no (CP–), they ($n = 29$) are classified as low risk. The remaining 27 patients are checked for whether any other symptom is present. If the answer is

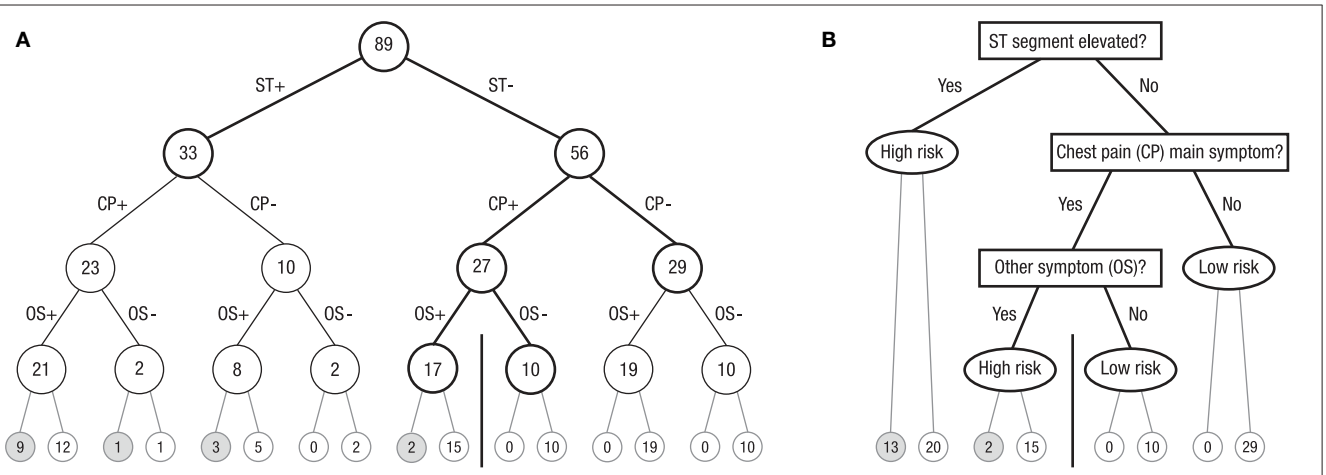

FIGURE 5 | (A) Full natural frequency tree for the Green and Mehr (1997) data on 89 patients with severe chest pain. The goal is to determine whether these patients are at high or low risk for heart disease. ST denotes a particular pattern in the electro cardiogram, CP denotes chest pain, OS denotes "at least one other symptom," "+" denotes present, and "–" denotes absent. Numbers in circles denote number of patients. **(B)** Fast-and-frugal classification tree obtained by pruning the natural frequency tree. The ranking of cues and the exit structure are determined by the ZigZag method (in the present case, ZigZag-val and ZigZag-sens, as explained in the text, lead to the same trees). Questions in rectangles specify which cues are looked up at this level for each of the patients in the corresponding circles in **(A)**. Depending on whether this cue value is positive or negative, either a new question is asked or the tree in **(B)** is exited and a classification decision is made (oval). The accuracy of these classification decisions is shown by the number of patients below these oval exit nodes: The number of patients who actually had a heart attack is displayed in the left of the two adjacent end nodes in the lowest layer, and the number of those who did not have one is displayed in the corresponding end node on the right. All patients to the left of the vertical bar in **Figure 1B** are classified as high risk, and all patients to its right are classified as low risk.

yes (OS+), they ($n = 17$) are classified as high risk; the others (OS–) are classified as low risk ($n = 10$). Each tree level corresponds to one cue, and the ranking of cues can follow simple heuristic procedures. Green and Mehr reported that diagnosis according to this fast-and-frugal tree was more accurate than both physicians' clinical judgment and logistic regression.

Two important features of the construction of a fast-and-frugal tree are the ranking of cues and its exit structure, that is, whether an exit is to the left or to the right (with the convention that branches defined by positive cue values will always be displayed at the left). One possible ranking, called the ZigZag-val method, is achieved by using the predictive values of the cues. The *positive predictive value* of a cue is the proportion of cases with a positive outcome among all cases with a positive cue value [i.e., $p(H|D)$] and the *negative predictive value* is the proportion of cases with a negative outcome among all cases with a negative cue value [i.e., $p(-H|-D)$]. The ZigZag-val tree has a left exit for levels 1 to k, where k is the smallest natural number so that $1/2^k$ is less than the ratio of the base rate of the disease divided by the base rate of healthy patients. For the levels after the kth level, the tree alternates between "yes" and "no" exits at each level, and a choice is made according to the cue with the greatest positive (for "yes") or negative (for "no") predictive value among the remaining cues (Martignon et al., 2008). A second method for tree construction, ZigZag-sens, has a left exit for levels 1 to k. For the levels after the kth level, the tree alternates between "yes" and "no" exits at each level, and a choice is made according to the cue with the greatest positive sensitivity [i.e., the greatest $p(D|H)$] or specificity [i.e., the greatest $p(-D|-H)$] among the remaining ones. Ties in the process are broken randomly.

Fast-and-frugal trees—those ranked according to positive and negative predictive value or according to sensitivity and specificity—radically reduce the complexity of full natural frequency trees. Their performance can be impressive. In the tree displayed in **Figure 5**, the lowest layer in **Figure 5A** displays the number of patients who after classification actually had a heart attack (left end nodes) and those who did not (the corresponding end nodes to the right). The vertical bar in the lowest layer can be seen as cutoff. The fast-and-frugal tree in **Figure 5B** is arranged so that all nodes to its left ($n = 50$) are classified as high risk (yielding 15 hits and 35 false alarms), and every one of the 39 cases to right of the bar are classified as low risk (yielding 0 misses and 39 correct rejections). In particular, fast-and-frugal trees ranked by sensitivity and specificity yield ROC curves with large areas underneath. Such properties are fundamental for medical doctors to reduce costly errors, in particular, misses (for ROC curves and fast-and-frugal trees, see Luan et al., 2011).

Another class of trees that reduce complexity is that based on CART (Breiman et al., 1984); these trees are simple in execution but often require complicated computations for their construction. To reduce complexity while maintaining the tenets of the Bayesian attitude, the strategy is to adopt the Naïve Bayes approach. Its simplification consists of assuming that cues are independent conditional on presence or absence of the disease, so that the probability of disease given cues can be estimated as the product of the conditional probabilities of disease given each one of the cues.

However, the tradition among practitioners has been to make use of classification strategies based on some type of regression. For binary classification, logistic regression is the standard model used by practitioners. When using logistic regression one assigns a value of 0 to the "low" state of H_k and a value of 1 to the "high" state of H_k. The logistic regression equation is:

$$\frac{p\,(D|H_1,\ldots,H_n)}{1-p\,(D|H_1,\ldots,H_n)} = e^{\beta_0+\sum_k \beta_k H_k} \tag{4}$$

where the parameters are typically estimated from data.

Laskey and Martignon (2014) compared the predictive accuracy of these five classification methods using 11 data sets taken from medical domains. When the models were constructed based on 90% of the data set, Naïve Bayes performed best, achieving 80% accuracy, while Logistic Regression achieved 79%. CART, like the ZigZag-val tree, achieved 74% accuracy, while the ZigZag-sens tree achieved 72% accuracy (note that in Laskey and Martignon, ZigZag-val is labeled ZigZag tree and ZigZag-sens was computed but not reported). When the models were constructed based on 50% of the data, CART, ZigZag-val, and ZigZag-sens performed at the same level as when being fitted to 90% of the data, whereas Logistic Regression and Naïve Bayes lost one percentage point each. Even more surprising, when the training set amounted to only 15% of the data set, ZigZag-val outperformed logistic regression and CART. In an uncertain world, where large numbers of correlations need to be estimated, fast-and-frugal trees can reduce estimation error and can have a competitive advantage over more complex strategies, in particular for small learning samples (Luan et al., 2011).

Predictive accuracy is not the only important criterion in medical diagnosis. It is often essential to make a diagnosis quickly or with limited diagnostic information. All in all, fast-and-frugal trees make it possible to act on limited information, and by reducing estimation error, they can perform competitively in situations entailing high complexity and uncertainty. They accomplish this by inverting natural frequency trees, so that the outcome (or hypothesis) is no longer displayed at the top of the tree (as in **Figure 1**) but at the lowest layer (as in **Figure 5A**). Subsequently, they can be pruned by cutting off branches, that is, by introducing an exit at every layer of the tree (as in **Figure 5B**).

Cue Merging

We will now discuss another way of reducing tree complexity, which amounts to merging multiple cues into one single cue. It has been studied in a probability learning task by Garcia-Retamero et al. (2007a) and Garcia-Retamero et al. (2007b). Participants had to make pair comparisons based on three cues, C_1, C_2, and C_3, with a validity (i.e., proportion of correct inferences) of 80, 60, and 60%, respectively. The cues were *not* independent. Specifically, although the cues C_2 and C_3 had a relatively low validity, they could be merged—by applying simple

Boolean algebra—into one cue that had a validity of 100%. For instance, if C_2 AND C_3 was present, then the alternative to which the cue pointed was correct in 100% of the cases (in two other conditions, we constructed environments in which merging two cues with the OR combination and the XOR combination created a new cue with a validity of 100% as well). Participants were not informed about this structure, but they were told that the three cues represent whether some drugs have been given to two patients. Their task was to predict which of two patients had the higher blood pressure. In these studies, the mental models of the participants were manipulated. In one condition, participants were informed that the three drugs operate in three different systems (hormonal, nervous, blood) and in the other condition that they operated within the same system. Those participants who had been told that the three drugs operated via different systems assumed independence and did *not* detect the hidden cue structure. By contrast, a majority of those participants who had been informed that the drugs operated via the same system could not safely exclude independence and *did* detect the structure. In a mouselab task, they immediately clicked C_2 and C_3, inspected both values, and only if the merged cue was not present did they request C_1 (note that they started with C_2 and C_3 even though each of these had a lower validity than C_1).

As this study demonstrates, participants assume independence by default but can detect dependencies if these exist. Such detection is easy with a natural frequency representation, which obviously can be constructed even in a probability learning task. Once participants have learned that cues can be merged, they treat this new cue as a single one, even though it is composed of two (similar to the term *bachelor*, which requires the presence of two features, male and unmarried). This empirical demonstration brings to mind Green and Mehr's (1997) fast-and-frugal tree, in which one of the nodes also contains a merged cue—in that case, an OR conjunction of five cues (labeled "other symptom"; **Figure 5B**).

The common denominator between fast-and-frugal trees and cue merging is that both can simplify the structure of a complex natural frequency tree. Both exploit certain structures of information (such as conditional dependence) and are "ecologically rational" if these structures are present. Constructing fast-and-frugal trees amounts to inverting complex natural frequency trees (with a hypothesis at the top layer) into simple classification trees (with data at the top) that implement one-reason decision making. Such trees perform well if some cues are so informative that less predictive cues no longer add substantial predictive value and can hence be ignored. Cue merging amounts to combining several cues into one; these merged cues can lead to better inferences than any of the single cues used separately. In general, fast-and-frugal heuristics—including fast-and-frugal trees and simple heuristics for pair comparison, with or without merged cues—are ecologically rational if they are adapted to the structure of information in the environment (Martignon and Hoffrage, 1999, 2002; Todd et al., 2012). Future research has to address the question of what the crucial variables (e.g., number of cues) are that trigger switching from being a Bayesian to being fast and frugal. For a first step in this direction, see Martignon and Krauss (2003), and for an exploration of Bayesian inferences as a function of task characteristics, see Hafenbrädl and Hoffrage (2015).

The Effect of Natural Frequencies Can Be Amplified by Visual Representations

In Study 2, we used natural frequencies to instruct participants how to reason the Bayesian way. In this context, we also presented the frequency tree to participants (see Appendix II in Supplementary Material). Such a tree supports any text in explaining natural frequencies through a visualization of the information structure relevant to solve a Bayesian inference task. But trees are not the only tool that can serve this function. Others are icon arrays, Euler diagrams, frequency grids, unit squares, and roulette wheel diagrams (for an overview see Binder et al., 2015; Mandel, 2015). Garcia-Retamero and Hoffrage (2013) demonstrated that patients' performance in a basic Bayesian inference task could be improved through a frequency grid whose effect is above and beyond that of natural frequency representation in the written text. The most common visualizations used in teaching statistics in schools, however, tend to be 2×2 tables and tree diagrams, both of which explicitly contain numbers. Note that these visual aids can make use of natural frequencies or probabilities and improve participants' performance when natural frequencies are used: In a study by Steckelberg et al. (2004), the beneficial effect of natural frequencies was about the same in both conditions. By contrast, tree diagrams and 2×2 tables using probabilities (or relative frequencies) do not improve participants' performance—yet are omnipresent in textbooks on probability theory (for an empirical study on the effect of these visualizations beyond pure format effects, see Binder et al., 2015).

With respect to visualization of Bayesian reasoning situations with two hypotheses and more than two cue values, both trees and tables can be easily extended to illustrate such situations (e.g., for three cue values, see the tree in **Figure 1B**, and imagine a 2×3 table). Likewise, a situation with more than two hypotheses and a dichotomous cue can easily be represented by a tree (e.g., for three hypotheses, see the tree in **Figure 1C**, and imagine a 3×2 table). However, situations with more than two cues appear to be easier to represent by trees (e.g., **Figure 1D**) than by tables. The ease of constructing and generalizing tree diagrams containing natural frequencies was the reason for choosing this visual aid in Study 2.

All in all, the available evidence shows that natural frequencies can facilitate Bayesian reasoning in situations of risk, that is, where probabilities are (assumed to be) known, as in textbook problems. The novel insights of this article are that this power extends to complex Bayesian tasks and that teaching natural frequencies in basic tasks generalizes to complex tasks. These insights correct the widespread claim that people are not built to reason the Bayesian way, and, more important, they provide an efficient tool to teach Bayesian reasoning even in complex situations.

Author Note

The studies reported in this manuscript were approved by the ethic committee of the Max Planck Institute for Human Development, Berlin, and were carried out with written informed consent from all participants. We would like to thank the three reviewers and the editors for their valuable feedback and Rona Unrau for editing the manuscript. This work was supported by grant 100014_140503 from the Swiss National Science Foundation.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01473

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Ayal, S., and Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgm. Decis. Mak.* 9, 226–242. Available online at: http://www.decisionsciencenews.com/sjdm/journal.sjdm.org/12/12714/jdm12714.pdf

Binder, K., Krauss, S., and Bruckmaier, G. (2015). Effects of visualizing statistical information—An empirical study on tree diagrams and 2 x 2 tables. *Front. Psychol.* 6:1186. doi: 10.3389/fpsyg.2015.01186

Bramwell, R., West, H., and Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: experimental study. *BMJ* 333, 284–286. doi: 10.1136/bmj.38884.663102.ae

Brase, G. L. (2002). Which statistical formats facilitate what decisions? The perception and influence of different statistical information formats. *J. Behav. Decis. Mak.* 15, 381–401. doi: 10.1002/bdm.421

Brase, G. L. (2007). The (in)flexibility of evolved frequency representations for statistical reasoning: Cognitive styles and brief prompts do not influence Bayesian inference. *Acta Psychol. Sin.* 39, 398–405.

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychon. Bull. Rev.* 15, 284–289. doi: 10.3758/PBR.15.2.284

Brase, G. L., and Hill, W. T. (2015). Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Front. Psychol.* 6:340. doi: 10.3389/fpsyg.2015.00340

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth.

Chapman, G. B., and Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgm. Decis. Mak.* 4, 34–40. Available online at: http://www.sjdm.org/journal/8708/jdm8708.pdf

Chater, N., and Oaksford, M. (eds.). (2008). *The Probabilistic Mind: Prospects for Bayesian Cognitive Science.* Oxford: Oxford University Press.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends Cogn. Sci.* 10, 335–344. doi: 10.1016/j.tics.2006.05.006

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

Daston, L. (1988). *Classical Probability in the Enlightenment*. Princeton, NJ: Princeton University Press.

Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine: problems and opportunities," in *Judgment under Uncertainty: Heuristics and Biases,* eds D. Kahneman, P. Slovic, and A. Tversky (Cambridge: Cambridge University Press), 249–267.

Edwards, W. (1968). "Conservatism in human information processing," in *Formal Representation of Human Judgment,* ed B. Kleinmuntz (New York, NY: Wiley), 17–52.

Gallistel, C. R. (1990). *The Organization of Learning.* Cambridge, MA: MIT Press.

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

Garcia-Retamero, R., Hoffrage, U., and Dieckmann, A. (2007a). When one cue is not enough: Combining fast and frugal heuristics with compound cue processing. *Q. J. Exp. Psychol.* 60, 1197–1215. doi: 10.1080/17470210600937528

Garcia-Retamero, R., Hoffrage, U., Dieckmann, A., and Ramos, M. M. (2007b). Compound cue processing within the fast and frugal heuristics approach in non-linearly separable environments. *Learn. Motiv.* 38, 16–34. doi: 10.1016/j.lmot.2006.05.001

Gigerenzer, G. (2002). *Calculated Risks: How to Know when Numbers Deceive You*. New York, NY: Simon and Schuster.

Gigerenzer, G. (2014). *Risk Savvy: How to Make Good Decisions*. New York, NY: Viking.

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Rev. Philos. Psychol.* 6, 361–383. doi: 10.1007/s13164-015-0248-1

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychol. Sci. Public Interest* 8, 53–96. doi: 10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G. (1996). On narrow norms and vague heuristics: a reply to Kahneman and Tversky. *Psychol. Rev.* 103, 592–596. doi: 10.1037/0033-295X.103.3.592

Gigerenzer, G., and Gray, J. A. M. (eds.). (2011). *Better Doctors, Better Patients, Better Decisions: Envisioning Healthcare in 2020*. Cambridge, MA: MIT Press.

Gigerenzer, G., Hertwig, R., and Pachur, T. (2011). *Heuristics: The Foundations of Adaptive Behavior*. New York, NY: Oxford University Press.

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., and Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: A reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev.* 106, 425–430. doi: 10.1037/0033-295X.106.2.425

Gigerenzer, G., and Hoffrage, U. (2007). The role of representation in Bayesian reasoning: correcting common misconceptions. *Behav. Brain Sci.* 30, 264–267. doi: 10.1017/S0140525X07001756

Gigerenzer, G., and Murray, D. J. (1987). *Cognition as Intuitive Statistics*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G., and Selten, R. (2001). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Gigerenzer, G., Todd, P. M., and The ABC Research Group. (1999). *Simple Heuristics that Make us Smart*. New York, NY: Oxford University Press.

Green, L., and Mehr, D. R. (1997). What alters physicians' decisions to admit to the coronary care unit? *J. Fam. Pract.* 45, 219–226.

Hafenbrädl, S., and Hoffrage, U. (2015). Towards an ecological analysis of Bayesian inference: How task characteristics influence responses. *Front. Psychol.* 6:939. doi: 10.3389/fpsyg.2015.00939

Harlow, H. F. (1958). The nature of love. *Am. Psychol.* 13, 573–685. doi: 10.1037/h0047884

Hasher, L., and Zacks, R. T. (1979). Automatic and effortful processes in memory. *J. Exp. Psychol.* 108, 356–388. doi: 10.1037/0096-3445.108.3.356

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychol. Sci.* 15, 534–539. doi: 10.1111/j.0956-7976.2004.00715.x

Hertwig, R., Hoffrage, U., and The ABC Research Group. (2013). *Simple Heuristics in a Social World.* New York, NY: Oxford University Press.

Hill, W. T., and Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *Q. J. Exp. Psychol.* 65, 2343–2368. doi: 10.1080/17470218.2012.687004

Hoffrage, U., and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Acad. Med.* 73, 538–540. doi: 10.1097/00001888-199805000-00024

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition* 84, 343–352. doi: 10.1016/S0010-0277(02)00050-1

Hoffrage, U., Hafenbrädl, S., and Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Front. Psychol.* 6:642. doi: 10.3389/fpsyg.2015.00642

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science* 290, 2261–2262. doi: 10.1126/science.290.5500.2261

Johnson, E. D., and Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Front. Psychol.* 6:938. doi: 10.3389/fpsyg.2015.00938

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Macmillan.

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Kleiter, G. D. (1994). "Natural sampling. Rationality without base rates," in *Contributions to Mathematical Psychology, Psychometrics, and Methodology,* eds G.H. Fisher and D. Laming (New York, NY: Springer), 375–388.

Kurzenhäuser, S., and Hoffrage, U. (2002). Teaching Bayesian reasoning: an evaluation of a classroom tutorial for medical students. *Med. Teach.* 24, 516–521. doi: 10.1080/0142159021000012540

Laskey, K., and Martignon, L. (2014). "Comparing fast and frugal trees and Bayesian networks for risk assessment," in *Proceedings of the 9th International Conference on Teaching Statistics*, ed K. Makar (Flagstaff, AZ: International Statistical Institute and International Association for Statistical Education). Available online at: http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8I4_LASKEY.pdf

Lindsey, S., Hertwig, R., and Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics* 6, 147–163. Available online at: (see http://pubman.mpdl.mpg.de/pubman/item/escidoc:2101705/component/escidoc:2101704/SL_Communicating_2003.pdf)

Luan, S., Schooler, L. J., and Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychol. Rev.* 118, 316–338.

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Front. Psychol.* 6:387. doi: 10.3389/fpsyg.2015.00387

Martignon, L. (2011). "Future Teachers' training in statistics: the situation in Germany," in *Teaching Statistics in School-Mathematics-Challenges for Teaching and Teacher Education: A Joint ICMI/IASE Study*, eds C. Batanero, G. Burrill, and C. Reading (Netherlands: Springer), 33–36. Available online at: http://link.springer.com/book/10.1007/978-94-007-1131-0

Martignon, L., and Hoffrage, U. (1999). "Why does one-reason decision making work? A case study in ecological rationality," in *Simple Heuristics that Make us Smart,* eds G. Gigerenzer, P. M. Todd, and the ABC Research Group (New York, NY: Oxford University Press), 119–140.

Martignon, L., and Hoffrage, U. (2002). Fast, frugal and fit: Simple heuristics for paired comparison. *Theory Decis.* 52, 29–71. doi: 10.1023/A:1015516217425

Martignon, L., Katsikopoulos, K. V., and Woike, J. K. (2008). Categorization with limited resources: a family of simple heuristics. *J. Mathematical Psychol.* 52, 352–361.

Martignon, L., and Krauss, S. (2003). "Can l'homme eclaire be fast and frugal? Reconciling Bayesianism and bounded rationality," in *Emerging Perspectives on Judgment and Decision Research,* eds S. L. Schneider and J. Shanteau (Cambridge: Cambridge University Press), 108–122.

Martignon, L., Vitouch, O., Takezawa, M., and Forster, M. R. (2003). "Naive and yet enlightened: From natural frequencies to fast and frugal decision trees," in *Thinking: Psychological Perspective on Reasoning, Judgment, and Decision Making,* eds D. K. Hardman and L. Macchi (New York, NY: Wiley), 189–211.

Massaro, D. (1998). *Perceiving Talking Faces*. Cambridge, MA: MIT Press.

McElreath, R., Wallin, A., and Fasolo, B. (2013). "The evolutionary rationality of social learning," in *Simple Heuristics in a Social World,* eds Hertwig, R., Hoffrage, U., and The ABC Research Group (New York, NY: Oxford University Press), 381–408.

Navarrete, G., and Mandel, D. (2015). Available online at: http://journal.frontiersin.org/researchtopic/2963/improving-bayesian-reasoning-what-works-and-why#overview

Pearce, J. M. (1997). *Animal Learning and Cognition: An Introduction*. Hove: Psychology Press.

Richerson, P. J., and Boyd, R. (2008). *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago, IL: University of Chicago Press.

Rosenbaum, S. E., Glenton, C., and Oxman, A. D. (2010). Summary-of-findings tables in Cochrane reviews improved understanding and rapid retrieval of key information. *J. Clin. Epidemiol.* 63, 620–626. doi: 10.1016/j.jclinepi.2009.12.014

Royal College of Obstetricians and Gynecologists. (2008). *Clinical Governance Advice No. 7*. Available online at: https://www.rcog.org.uk/globalassets/documents/guidelines/clinical-governance-advice/cga7-15072010.pdf (Accessed August 14, 2015).

Ruscio, J. (2003). Comparing Bayes' theorem to frequency-based approaches to teaching Bayesian reasoning. *Teach. Psychol.* 30, 325–328.

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Sirota, M., Juanchich, M., and Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychon. Bull. Rev.* 21, 198–204. doi: 10.3758/s13423-013-0464-6

Sirota, M., Kostovièová, L., and Vallée-Tourangeau, F. (2015a). Now you Bayes, now you don't: effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychon. Bull. Rev.* 22, 1465–1473. doi: 10.3758/s13423-015-0810-y

Sirota, M., Vallée-Tourangeau, G., Vallée-Tourangeau, F., and Juanchich, M. (2015b). On Bayesian problem-solving: helping Bayesians solve simple Bayesian word problems. *Front. Psychol.* 6:1141. doi: 10.3389/fpsyg.2015.01141

Spiegelhalter, D., and Gage, J. (2014). "What can education learn from real-world communication of risk and uncertainty," in *Sustainability in Statistics Education. Proceedings of the 9th International Conference on Teaching Statistics (ICOTS9, July,* 2014*),* eds K. Makar, B. de Sousa, and R. Gould (Flagstaff, AZ; Voorburg: International Statistical Institute. Available online at: http://icots.info/9/proceedings/pdfs/ICOTS9_PL2_SPIEGELHALTER.pdf (Accessed August 23, 2015).

Steckelberg, A., Balgenorth, A., Berger, J., and Mühlhauser, I. (2004). Explaining computation of predictive values: 2×2 table versus frequency tree. A randomized controlled trial [ISRCTN74278823]. *BMC Med. Educ.* 4:13. doi: 10.1186/1472-6920-4-13

Thaler, R. H., and Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Till, C. (2013). "Risk literacy: first steps in primary school," in *Sustainability in Statistics Education. Proceedings of the 9th International Conference on Teaching Statistics (ICOTS9),* eds K. Makar, B. de Sousa, and R. Gould (Voorburg: International Statistical Institute).

Todd, P. M., Gigerenzer, G., and the, A. B. C., Research Group (2012). *Ecological Rationality: Intelligence in the World*. New York, NY: Oxford University Press.

Trevena, L., Zikmund-Fisher, B., Edwards, A., Gaissmaier, W., Galesic, M., Han, P., et al. (2012). "Presenting probabilities," in Update of the International Patient Decision Aids Standards (IPDAS) Collaboration's Background Document. Chapter C, eds R. Volk and H. Llewellyn-Thomas. Available online at: http://ipdas.ohri.ca/resources.html (Accessed August 14, 2015).

Tubau, E. (2008). Enhancing probabilistic reasoning: the role of causal graphs, statistical format and numerical skills. *Learn. Individ. Differ.* 18, 187–196. doi: 10.1016/j.lindif.2007.08.006

Waldmann, M. R., and Martignon, L. (1998). "A Bayesian network model of causal learning," in *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, eds M. A. Gernsbacher and S. J. Derry (Mahwah, NJ: Erlbaum), 1102–1107.

Woloshin, S., and Schwartz, L. M. (2011). Communicating data about the benefits and harms of treatment: a randomized trial. *Ann. Intern. Med.* 155, 87–97. doi: 10.7326/0003-4819-155-2-201107190-00004

Zhu, L., and Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition* 98, 287–308. doi: 10.1016/j.cognition.2004.12.003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 April 2015
doi: 10.3389/fpsyg.2015.00387

Instruction in information structuring improves Bayesian judgment in intelligence analysts

David R. Mandel *

Socio-Cognitive Systems Section, Defence Research and Development Canada and Department of Psychology, York University, Toronto, ON, Canada

An experiment was conducted to test the effectiveness of brief instruction in information structuring (i.e., representing and integrating information) for improving the coherence of probability judgments and binary choices among intelligence analysts. Forty-three analysts were presented with comparable sets of Bayesian judgment problems before and immediately after instruction. After instruction, analysts' probability judgments were more coherent (i.e., more additive and compliant with Bayes theorem). Instruction also improved the coherence of binary choices regarding category membership: after instruction, subjects were more likely to invariably choose the category to which they assigned the higher probability of a target's membership. The research provides a rare example of evidence-based validation of effectiveness in instruction to improve the statistical assessment skills of intelligence analysts. Such instruction could also be used to improve the assessment quality of other types of experts who are required to integrate statistical information or make probabilistic assessments.

Keywords: instructional methods, Bayesian judgment, probability judgment, information structuring, coherence

OPEN ACCESS

Edited by:
Bernhard Hommel, Leiden University, Netherlands

Reviewed by:
Gary L. Brase, Kansas State University, USA
Jean Baratgin, Université Paris 8, France

***Correspondence:**
David R. Mandel, Socio-Cognitive Systems Section, Defence Research and Development Canada, 1133 Sheppard Avenue West, Toronto, ON M3K 2C9, Canada
david.mandel@drdc-rddc.gc.ca

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *09 February 2015*
Accepted: *18 March 2015*
Published: *08 April 2015*

Citation:
Mandel DR (2015) Instruction in information structuring improves Bayesian judgment in intelligence analysts. Front. Psychol. 6:387. doi: 10.3389/fpsyg.2015.00387

Introduction

Categorization under uncertainty is a basic fact of life. In a wide range of contexts, both personal and professional, people strive to accurately categorize "objects," including, at times, themselves. Yet in many, if not most, cases, the correct category to which an object belongs is not immediately apparent. Instead, one might have to generate hypotheses about putative category membership. Moreover, the evidence one has at one's disposal is usually inconclusive, serving at best to amplify or attenuate support for the hypotheses under consideration. In other words, the evidence may not fully eliminate uncertainty about category membership yielding a definitive answer. Indeed, it is primarily because most everyday judgment and reasoning is made under conditions of uncertainty that the dominant normative paradigm for assessing reasoning quality has shifted from a truth functional logic of certain deduction to a Bayesian logic of uncertain deduction (e.g., Oaksford and Chater, 2007; Evans, 2012; Baratgin et al., 2014).

The literature on Bayesian reasoning is rich and the focus of this paper is restricted to two aspects of it: Bayes theorem and the complementarity constraint (Baratgin and Noveck, 2000), which is a special case of the axiom of finite additivity of closed subsets, often called the additivity principle in cognitive psychology (e.g., Tversky and Koehler, 1994; Villejoubert and Mandel, 2002). The paper does not, for instance, address aspects of Bayesian reasoning having to do with the alternative logical and subjectivist stances on Bayesianism, nor does it examine adherence to the dynamic coherence criterion

known as the conditioning principle (for an overview of these other issues, see Baratgin and Politzer, 2006). Rather, the aspects addressed here pertain to static coherence criteria reflecting the normative view that probability is additive (Kolmogorov, 1950). Finally, although my focus is on the aforementioned aspects of Bayesianism, I neither presume nor wish to suggest that Bayesian approaches are the only viable normative frameworks for reaching probabilistic inferences under conditions of uncertainty (e.g., Lewis, 1976; Thagard, 1989; Douven and Schupbach, 2015). Indeed, as few others have noted (e.g., see Walliser and Zwirn, 2002; Baratgin and Politzer, 2006, 2010), Bayesian revision is normative in a restricted set of problem representations known as *focusing* cases—namely, cases where the original set of possible worlds is preserved rather than transformed over time. This is the type of problem studied in the present research, where only two categories exist and new information cannot invalidate either category. However, in many other cases (e.g., see Baratgin, 2009; Cozic, 2011) new information may transform the set of categories (or hypotheses) being considered. In such *updating* cases, Lewis's (1976) imaging rule provides a normative solution for probability redistribution.

For our purposes, let Ω represent an event space comprised of elementary events, w_i, that is partitioned into a non-empty, closed family of subsets A. The focus in this paper is specifically on subset families that exhibit binary complementarity; namely, in which $\{A, B\} \in$ A, $A \cap B = \emptyset$ (i.e., A and B are mutually exclusive), A $\equiv A \cup B$ (i.e., A and B exhaustively partition A). Indeed, since $A \Leftrightarrow \neg B$ (and likewise $B \Leftrightarrow \neg A$), let us use $\neg A$ instead of B to remind ourselves that the two subsets are binary complements. For our purposes, let H_A and $H_{\neg A}$ represent mutually exclusive and exhaustive hypotheses about the category membership of a focal elementary event, w, which in subsequent examples given in this paper is a person whose category membership is unknown. Thus, H_A and $H_{\neg A}$ stand for the propositions that $w \in A$ and $w \in \neg A$, respectively. In the Bayesian context, the probabilities assigned to these complementary hypotheses may be revised in light of new evidence or data, D. These "posterior" probabilities (see Mandel, 2014a, for an explanation of the scare quotes), $P(H_A|D)$ and $P(H_{\neg A}|D)$, are the focus of most studies of Bayesian judgment, as they are in this paper.

Given the preceding definitions, the additivity principle for binary complements states that $P(H_A|D \cup H_{\neg A}|D) = P(H_A|D) + P(H_{\neg A}|D)$, where P stands for probability, a non-negative real number in the [0, 1] interval. Let $T = P(H_A|D) + P(H_{\neg A}|D)$. The complementarity constraint states that $T = 1$. In this paper, I break with the majority of papers that have followed Tversky and Koehler (1994) by calling normative violations in which $T < 1$ superadditive and violations in which $T > 1$ subadditive—terms which appear to mean precisely the opposite of what they are intended to convey. Instead, following Baratgin and Noveck (2000), I refer to cases where $T < 1$ as *subadditive* and to cases where $T > 1$ as *superadditive*. This properly places the emphasis on the additivity of the binary complements relative to unity rather than the other way around, and it is likely to be intuitive to readers outside this specific niche.

With some exceptions (e.g., Wallsten et al., 1993; Rottenstreich and Tversky, 1997; Juslin et al., 2003; see Mandel, 2005, for an explanation of differences obtained across studies), most studies have shown that people assign subadditive probabilities to binary complements (Macchi et al., 1999; Baratgin and Noveck, 2000; Windschitl et al., 2003, Experiment 4; Sloman et al., 2004; Mandel, 2005; Williams and Mandel, 2007; Mandel, 2008, Experiments 5 and 6). Additivity violations have also been shown to be systematic, following the non-normative tendency to judge $P(H_A|D)$ and $P(H_{\neg A}|D)$ on the basis of their inverse probabilities, $P(D|H_A)$ and $P(D|H_{\neg A})$, respectively (Villejoubert and Mandel, 2002). This tendency has been variably called the Fisherian algorithm (Gigerenzer and Hoffrage, 1995), the confusion hypothesis (Macchi, 1995), the conversion error (Wolfe, 1995), and the inverse fallacy (Koehler, 1996). Thus, if we let $T' = P(D|H_A)$ and $P(D|H_{\neg A})$, what Villejoubert and Mandel (2002) found was that subjects' T-values tracked the objective T' values such that they were subadditive when $T' < 1$ and superadditive when $T' > 1$.

The second coherence constraint of interest in this paper is Bayes theorem, which is a corollary of the rule of compound probabilities, $P(H_A \cap D) = P(D|H_A)P(H_A) = P(H_A|D)P(D)$. Bayes theorem can be expressed in various ways. The most common format discussed in the literature on Bayesian reasoning performance is Bayes identity, which in general form may be expressed,

$$P(H_i \mid D) = \frac{P(H_i)\,P(D|H_i)}{P(D)} = \frac{P(H_i)\,P(D|H_i)}{\sum_i P(H_i)\,P(D|H_i)}. \quad (1)$$

In the case of binary complements, using the terms defined earlier, we can express Bayes identity as

$$\begin{aligned} P(H_A \mid D) &= \frac{P(H_A)\,P(D|H_A)}{P(D)} \\ &= \frac{P(H_A)\,P(D|H_A)}{P(H_A)\,P(D \mid H_A) + P(H_{\neg A})\,P(D|H_{\neg A})}. \end{aligned} \quad (2)$$

However, as the rule of compound probability makes clear, Bayes theorem can also be expressed,

$$P(H_A \mid D) = \frac{P(H_A \cap D)}{P(D)} = \frac{P(H_A \cap D)}{P(H_A \cap D) + P(H_{\neg A} \cap D)}. \quad (3)$$

When people are asked to judge $P(H_A|D)$ on the basis of information sources such as $P(H_A)$—the base rate—and $P(D|H_A)$ and $P(D|H_{\neg A})$—sometimes referred to as "diagnostic" probabilities, only a minority cohere in their judgments with Bayes theorem (e.g., Kahneman and Tversky, 1972, 1973; Lyon and Slovic, 1976; Casscells et al., 1978; Villejoubert and Mandel, 2002). For example, consider the following problem:

> The probability of breast cancer is 1% for a woman at age 40 who participates in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

Using Bayes theorem, the probability that the woman has breast cancer given her test result is nearly 8%, yet Eddy (1982) found that 95 out of 100 physicians presented with the problem roughly an order of magnitude higher and similar results with physician or medical counselor samples have been found in other studies (Gigerenzer et al., 1998; Hoffrage and Gigerenzer, 1998; Garcia-Retamero and Hoffrage, 2013).

A ubiquitous explanation for the well-documented divergence between people's probability judgments and those computed on the basis of Bayes theorem is that people neglect, or at least underweight, base-rate information (Kahneman and Tversky, 1972, 1973; Lyon and Slovic, 1976; Bar-Hillel, 1980). However, without undermining the claim that base-rates are often underutilized, there is also reason to believe that the divergences reported may be due to the inverse fallacy discussed earlier (Eddy, 1982; Hamm, 1993; Koehler, 1996). For example, Villejoubert and Mandel (2002) kept base rates for two mutually exclusive and exhaustive categories equiprobable and invariant across a set of Bayesian reasoning problems. They found that most subjects judged probabilities in violation of Bayes theorem even though the possibility of base-rate underutilization was eliminated in their experiment. Moreover, the direction and magnitude of the mean difference between subjects' judgments and the Bayesian values tracked the value of the inverse probabilities, just as additivity violations had tracked the sum of the inverse probabilities[1]. As well, information search in Bayesian tasks focuses significantly more on the inverse probability of a focal hypothesis ($P(D|H_A)$) than on either the contrapositive conditional probability ($P(D|H_{\neg A})$) or the base-rate ($P(H_A)$), and the more subjects focused on the inverse probability, the less they focused on the base rate (Wolfe, 1995). Thus, base-rate neglect may be due in part to the inverse fallacy. Finally, even in cases where base-rate neglect has been invoked as an explanation of non-conformity with Bayes theorem, such as Eddy (1982) results for the mammography problem described earlier, the inverse fallacy better accounts for the aggregate findings (Mandel, 2014a).

Improving the Coherence of Probability Judgments

The literature reviewed earlier shows that people often do not conform to two important coherence constraints on probability judgment when given statistical information as input to their judgment process: they systematically deviate from both the complementarity constraint and Bayes theorem. These manifestations of incoherence are particularly troubling when made by professionals whose judgments may, in turn, provide input to consequential decision-making. Much attention, as already noted, has been devoted to normative violations of probability judgment committed by medical professionals.

Another group of experts who make probabilistic judgments are intelligence analysts. Intelligence analysis plays a vital role in national and international security, serving as key sources of information for a wide range of decision-makers including state leaders, policy makers, and military commanders. Despite the importance of intelligence analysis—and the centrality of probabilistic judgment in intelligence products (Kent, 1964; Zlotnick, 1972; Friedman and Zeckhauser, 2012), there are few behavioral studies of analytical judgment quality (Pool, 2010). Probabilistic assessments underlie virtually all forecasts made by intelligence agencies. Moreover, intelligence analysts, managers, and trainers acknowledge that the predictive function of intelligence is roughly as important as the narrative descriptive function (Adams et al., 2012). Although one study has found that strategic intelligence forecasts showed good discrimination and calibration (Mandel and Barnes, 2014), the extent to which analytical judgments are coherent has not been addressed in an intelligence analyst sample. Such research is needed because intelligence analysts must often revise their hypotheses and beliefs based on missing and uncertain evidence.

Nevertheless, few, if any, analysts receive training in probabilistic belief revision. More commonly, analysts receive brief training lessons that highlight the "mindsets and biases" to which all humans are prone. In such training, analysts are taught, for instance, to "beware of overconfidence" and to "avoid confirmation bias," but they are not routinely taught how to assess their own or others' coherence or accuracy. Few of the structured analytic techniques that analysts may use to support their assessments have been scientifically tested (Pool, 2010). Most are based on what made sense to their developers, most of whom do not have backgrounds in behavioral science. Moreover, members of the intelligence community have identified the need for evidence-based research on analytical processes that support effectiveness as a priority (Adams et al., 2012). One aim of the present research was to examine the extent to which intelligence analysts' probability judgments conform to the complementarity constraint and Bayes theorem in statistical integration tasks like the mammography problem. And a second aim was to test whether brief instruction in information structuring would have a positive effect on the quality of intelligence analysts' probability judgments. In that regard, the present research represents a rare test of the effectiveness of instruction that could be used to improve intelligence analysts' probabilistic reasoning skill.

The present research leverages recent developments in improving Bayesian reasoning. It is well established that a greater proportion of subjects in Bayesian reasoning studies provide Bayesian answers or describe a Bayesian computational process when the information provided to them is expressed in terms of natural frequencies (Gigerenzer and Hoffrage, 1995; Cosmides and Tooby, 1996; also see Kleiter, 1994). To express in natural frequencies information such as that given in the mammography problem, one would begin with a hypothetical reference class that could be easily broken down into subsamples. For instance, one might start with 1000 women aged 40 who participate in routine screening. The 1% base-rate would then be represented by subsets of 10 women who have breast cancer (H_A) and 990 who do not ($H_{\neg A}$). The former subset is further decomposed into true-positive ($H_A \cap D_+$, where D_+ stands for the positive-test result) and false-negative ($H_A \cap D_-$, where D_- stands for the negative-test result that was not obtained) subsets (8 and 2 cases, respectively), and the latter is likewise decomposed into true-negative

[1]This is by necessity: if $T < 1$, then the mean bias (i.e., the mean deviation between the subject's posterior probability and the values given by Bayes theorem) must be negative, representing underestimation, by the same degree. Likewise when $T > 1$; then, mean bias must represent overestimation to the same degree.

($H_{\neg A} \cap D_-$) and false-positive ($H_{\neg A} \cap D_+$) subsets (895 and 95 cases, respectively). When the information is represented as such, it is easier to calculate the "short form" of Bayes theorem shown in Equation 3. The numerator of this equation is already identified ($f(H_A \cap D_+) = 8$) and the denominator simply involves adding the two subsets containing D_+ (i.e., $8 + 95 = 103$). Even without dividing, one might appreciate that the value 8/103 is slightly less than 8%.

Although the finding that restructuring of statistical information, such as that given in the mammography problem, into the natural frequency format just described yields better correspondence to Bayes theorem, the bases for the effect are the subject of much debate. Given that the present research does not focus on that "why" question, but rather uses the descriptive findings to explore whether Bayesian reasoning may be improved through instruction, I merely note that it is important to separate the descriptive findings from the theoretical accounts of them that have been proposed. As well, most adaptationists (e.g., see Gigerenzer and Hoffrage, 2007) and dual-systems theorists (Barbey and Sloman, 2007) do not strongly disagree that the beneficial effect of natural frequency formats derive from a combination of factors, including clarifying nested set structure of the relevant statistical data, improve the compatibility between evidence and queries, and reduce the computational complexity of task at hand (Mandel, 2007; Ayal and Beyth-Marom, 2014). More importantly, for the present purposes, most researchers agree that natural frequency presentations of statistical information in Bayesian reasoning tasks tend to facilitate Bayesian reasoning and improve Bayesian judgment.

The use of natural frequencies to convey probabilistic evidence is further augmented by the use of visual representations that reinforce the nested-set structure of diagnostic and base-rate evidence (Cosmides and Tooby, 1996). Indeed, visual representations can facilitate Bayesian reasoning by clarifying nested-set relations even when natural frequencies are not explicitly encoded in the representations (Sloman et al., 2003; Sirota et al., 2015). Such representations can also clarify the logical relations and the structure of arguments in support of alternative normative views on belief revision tasks (Mandel, 2014b). However, in at least some studies, visual representations that encode natural frequency information directly through icons or numerical values have been shown to be more effective than visualizations that clarify set structure but do not explicitly encode the frequency data, such as Euler diagrams (Sedlmeier, 1999, chapter 6; Brase, 2008, 2014). Although not all studies have shown such an advantage (e.g., Sirota et al., 2015), no study has reported the opposite effect; namely, better performance with nested-set representations that do not include explicit frequency encoding than with nested-set representations that do include such coding.

The use of visual representations of natural frequencies has also been shown to be an effective instructional method for improving compliance with Bayes theorem. Sedlmeier and Gigerenzer (2001; see also Sedlmeier, 1999) found that a single 1–2 h session of practice-based instruction in Bayesian reasoning facilitated performance on Bayesian judgment tasks. The performance boost immediately after instruction was large regardless of whether the instruction used rule-based training in the application of Bayes theorem or whether it used a natural sampling representation such as a frequency grid or frequency tree. The long-term effect of instruction, however, showed a clear advantage for instruction that relied on a natural sampling representation of the information provided in a given problem. In three experiments, on average, subjects who received such instruction performed as well at the longest-term test phase (i.e., 5 weeks in two experiments and 3 months in another experiment) as they did in the immediate test phase. In contrast, rule-based instruction showed substantial decrements by the last test phases in all experiments.[2] The instructional benefit of frequency-based visual representations on Bayesian reasoning has been confirmed in other studies as well (Kurzenhäuser and Hoffrage, 2002; Ruscio, 2003; McCloy et al., 2007).

The present research examined the effect of instruction in information structuring on adherence to the complementarity constraint and Bayes theorem in a sample of intelligence analysts who were undergoing military intelligence training. Unlike earlier studies of instruction effects on Bayesian judgment (e.g., Sedlmeier and Gigerenzer, 2001; McCloy et al., 2007; Sirota et al., 2015), the aim of this research was not to compare different modes of instruction. Rather, the effect of a single instructional mode using a natural sampling approach with natural-frequency-tree diagrams was examined, given that this mode has already been shown to yield stable long-term improvement in conditional probability judgment. Unlike earlier research on instruction, however, this research used a pre-post design to assess the effect of instruction on complementarity constraint violations and deviations from Bayes theorem. The vast majority of studies of Bayesian reasoning have used problems with binary outcome categories corresponding to H_A and $H_{\neg A}$ but have only queried subjects about one of the two hypotheses, H_A. Thus, they were unable to examine the effect of Bayesian instruction on the additivity of subjects' judgments.

Moreover, the study was designed so that predictions regarding the direction of error could be made on the basis of the inverse fallacy, which, as noted earlier, has successfully accounted for both additivity violations and deviations from Bayes theorem (Villejoubert and Mandel, 2002). Specifically, assuming that the grand mean of T across subjects, hypotheses, and test items is additive, it was predicted that $\overline{T} < 1$ if $T' < 1$ and that $\overline{T} > 1$ if $T' > 1$. Naturally, if there were to be an overall bias toward a form of nonadditivity, the predictions would be relaxed, taking the form of the mean difference prediction $\overline{T}|(T' < 1) < \overline{T}|(T' > 1)$. That is, a general bias in additivity would negate the predicted reflection around additivity. Given that most studies of adherence to the complementarity constraint have reported subadditivity, this form of nonadditivity is the likelier candidate. Indeed, Williams and Mandel (2007) found subadditivity for conditional probability judgments of binary complements. Although Villejoubert and Mandel (2002) did not report whether there was an overall bias in T, it is evident by averaging the mean T-values in the last column of **Table 2** in that paper that the grand mean

[2]The one exception was in Study 1b of Sedlmeier and Gigerenzer (2001) where subjects were incentivized through bonuses and where rule-based and natural sampling methods yielded comparable performance.

(where the simple means were elicited within subjects) is equal to 0.916, a value that reflects subadditivity. Given that the numerical characteristics of the test items used in the present research were drawn from Villejoubert and Mandel (2002), there is good reason to expect an overall bias toward subadditivity.

Finally, an aim of the research was to examine the coherence between subjects' probability judgments and their binary forced choice of the target's category membership. Presumably, subjects would choose the category to which they assigned a higher probability. However, studies of Bayesian judgments have not asked subjects to make a discrete choice in addition to making their probability judgments. Thus, it is of interest to verify whether, in fact, subjects do invariably choose in accordance with the higher assigned probability. And, to the extent that they do not, it is of interest to examine whether instruction might attenuate this form of incoherence. Since judgments are often a precursor to decisions and actions, this is a question that is of more that academic interest.

Materials and Methods

Subjects

Forty-three intelligence analyst trainees participated in the research during regular course time at the Canadian Forces School for Military Intelligence at Canadian Forces Base Kingston in Kingston Ontario, Canada. Twelve trainees were from a senior analysts' course, 16 were from an intermediate, basic intelligence officers' course and 15 were from a junior course. The entry requirements were an undergraduate degree for the intermediate course and completion of Grade 10 high school for the junior course. Trainees in the senior course had to have successfully completed the intermediate course. Demographic information was not recorded. However, over 90% of subjects were male. Subjects were informed that their participation was voluntary and that they would not be remunerated for their time. No student refused to participate.

Procedure

Subjects were introduced to the study in class by being told that intelligence analysts are routinely called upon to make assessments under conditions of uncertainty, where the information they receive may be probabilistic in nature. Subjects were further told that analysts must often revise their beliefs about hypotheses or events on the basis of new, but once again, uncertain information. After this preliminary statement, subjects were informed that they had the opportunity to participate in research aimed at improving their judgment abilities. After consenting to participate, subjects were given a pre-instruction booklet that contained eight probability judgment problems, described in detail below. Participants worked on the problems individually at their desks. The task was not strictly timed. However, subjects were told that they would have approximately 15 min to complete the task. All subjects completed the task in the allotted time. An anonymous subject code was generated by the subject and written on the pre-instruction booklet before it was returned to the experimenter so that it could be matched to the post-instruction booklet.

After returning the pre-instruction booklets, the experimenter told subjects that they would now be given a brief tutorial on how to accurately integrate different sources of probabilistic information to arrive at their own probabilistic assessments of different hypotheses that one might wish to test. The first run of this experiment was conducted on the senior course and the tutorial included a series of medical diagnosis examples. The second and third runs in the other courses used an alternative version of the tutorial, which was deemed by the senior instructor at the Canadian Forces School for Military Intelligence to be more relevant to the intelligence and security context, and which focused on detecting whether a human target was an insurgent. The two versions, however, had the same structure, length, and relevant content, differing only in terms of the domain of examples (i.e., medical diagnosis vs. intelligence target detection). Both versions of the full tutorial are presented in the Supplementary Materials.

The tutorial began with an example that presents the base-rate of a focal hypothesis, $P(H_A)$, and diagnostic probabilities, $P(D_+|H_A)$ and $P(D_+|H_{\neg A})$, where D_+ stands for data indicating a positive result on a diagnostic test. Subjects were asked how they might use that information to assess the conditional probability, $P(H_A|D_+)$—namely, the probability that the focal hypothesis was true given the data indicating a positive test result.

After being presented with the initial assessment task, subjects were asked to think about how they would go about making the assessment and to record their assessment. Next, the experimenter showed subjects how they could systematically work through the problem. Slides 3–5 in the tutorials were designed to show subjects how they could represent the information given to them as a natural-sampling-tree diagram. As each slide was presented, the experimenter read the textual content and pointed to the appropriate part of the diagram. Subjects were able to see the slides on a large projection screen located at the front of the classroom as well as on personal computer screens located directly in front of them on their desk spaces. On Slides 6–7, the experimenter worked through the solution, showing subjects how the information represented in the diagram could be arranged to answer the relevant question. The tutorial advises trainees to first identify the relevant set of cases that correspond to the condition, D_+, specified in the conditional probability, $P(H_A|D_+)$. Then, trainees are directed to identify the subset of those cases that conforms to the hypothesis—namely, $f(H_A \cap D_+)$. The corresponding diagrams made these points salient by color-coding the relevant sets of cases. The solution shown on Slide 6 represented those color-coded sets as an equation corresponding to the short form of Bayes theorem (Equation 3).

After being presented with the solution, subjects were asked to reflect on how it compared to their initial assessment (see Slide 7). Although this comparison was for pedagogical purposes, it is worth noting that many subjects commented that their estimates deviated from the correct value, and some confessed to not knowing how to integrate the information supplied (reinforcing Juslin, 2015, claim that while estimation may be very good, integration often falters).

After answering any questions subjects may have had, the experimenter moved onto the second example, which used the same cover story but asked subjects to imagine that the test result

had been negative (D_-) instead of positive. Subjects were asked to consider how they would assess the probability that the hypothesis was true given the negative test result, $P(H_A|D_-)$. After subjects gave their initial assessment, the experimenter worked through the problem in the same way as before, after which subjects compared their answers to the correct solution (see Slides 10–14).

The third example served to further illustrate that the approach taught could be used to answer other related questions, including questions framed in complementary ways (see Slide 15). Thus, whereas the second example asked subjects to assess $P(H_A|D_-)$, the third asked them to assess the probability of the alternative hypothesis given the same negative test result, $P(H_{\neg A}|D_-)$. Once again, the solution was presented using a natural-sampling-tree diagram (see Slides 16–17). However, subjects' attention was also drawn to the fact that the answers to the two last problems summed to 100%, and they were informed that this was no coincidence. **Figure 1** shows the natural-sampling-tree diagram with solutions to $P(H_A|D_-)$ on the left and $P(H_{\neg A}|D_-)$ on the right for the intelligence version of the tutorial.

On the next slide (Slide 18), the implicit lesson about the complementarity constraint just conveyed was made explicit. Subjects were introduced to the additivity principle and told that violations of additivity represented a form of incoherence in probability assessment. The tutorial then concluded with a summary of the following key points (see Slides 19–22): first, try to visually represent the information provided, such as in the natural-sampling-tree diagrams used in the tutorial; second, in preparation for information integration, think about the probability being assessed as a ratio and identify the relevant subsets that comprise the numerator and denominator, starting with the denominator because the numerator is always a subset of the denominator; and, finally, do the arithmetic required to produce the estimate.

After answering any questions subjects may have had, the experimenter administered the post-instruction booklet to subjects, which had an alternative set of problems much like the pre-training set (detailed in subsection Judgment Tasks). Once again, subjects were given approximately 15 min to complete the set of problems and they completed the task in the allotted time. When the booklets were returned, subjects were thanked, orally debriefed, and the experiment concluded.

Judgment Tasks

The primary judgment task assigned to subjects before and after instruction was adapted from that used by Villejoubert and Mandel (2002). The pre- and post-instruction booklets are included in the Supplementary Materials.

To summarize the task, subjects were asked to imagine that they were contestants on a game show who would be asked a series of skill-testing questions. They were to meet eight "mystery people" and, for each one, they would learn, following a query from the game-show host to the mystery person, whether a particular attribute (e.g., being a smoker) was present (D_+) or absent (D_-) in the individual. Half of the mystery people possessed the relevant attribute and the other half did not.

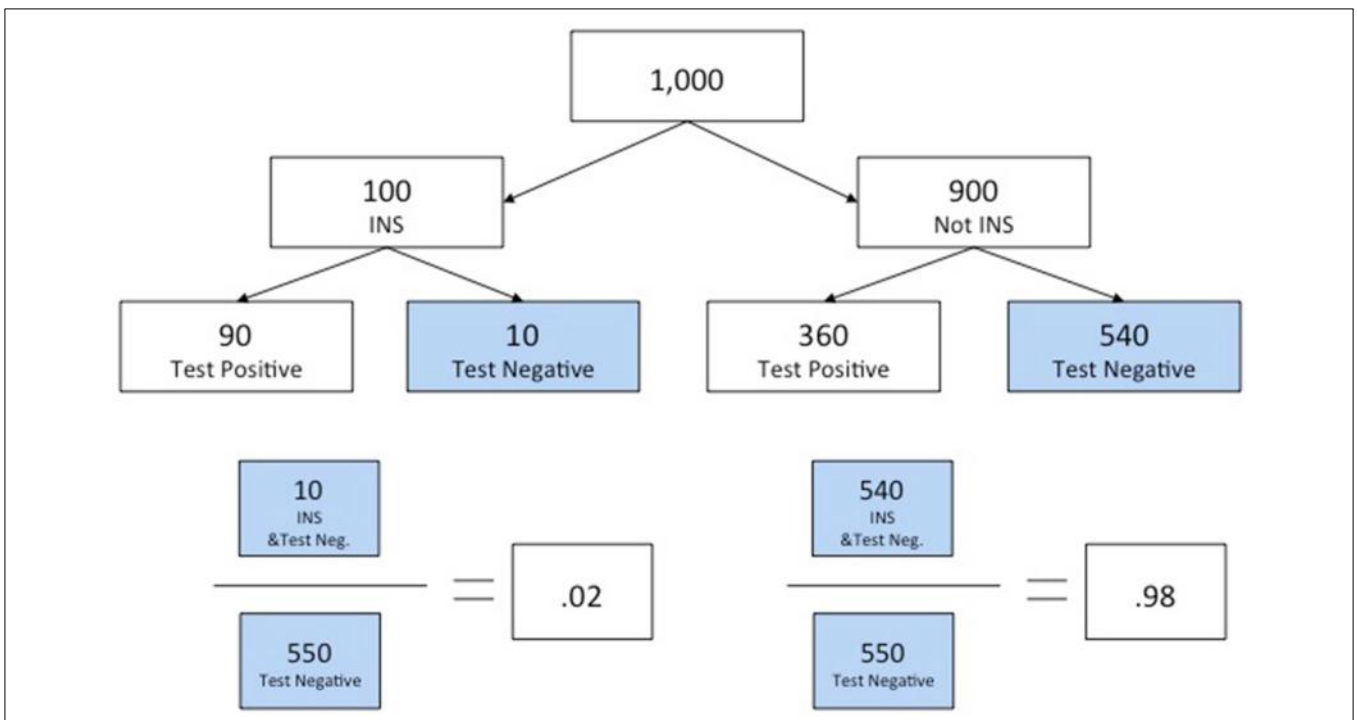

FIGURE 1 | Example from tutorial showing use of a natural sampling tree and providing solutions for assessments of alternative hypotheses defined by mutually exclusive and exhaustive subsets.

Subjects' task was to probabilistically assess the mystery person's group membership. Each person either belonged to Group A or Group B. For continuity with the prior discussion, let H_A stand for the hypothesis that the target person is a member of Group A and let $H_{\neg A}$ stand for the mutually exclusive, alternative hypothesis that the target person is a member of Group B. Subjects were informed that the overall population from which the sample of eight were said to be drawn was evenly divided and, thus, $P(H_A) = P(H_{\neg A}) = 0.5$. For each of the eight "encounters," subjects also learned the diagnostic probabilities of the attribute, $P(D_+|\mathrm{H}_A)$ and $P(D_+|H_{\neg A})$. Subjects were asked to estimate the probability that the target person was a member of Group A and then to estimate the probability that the person was a member of Group B on a "percentage chance" scale ranging from 0 (*absolutely no chance at all*) to 100 (*absolutely certain*) by writing a numerical value in a space provided. After giving their estimates, they were asked to make a binary choice regarding whether they thought the relevant mystery person was a member of Group A or Group B by circling one of the two options.

The diagnostic probabilities for the eight attributes (one per mystery person) are summarized in **Table 1**. Note that the pre- and post-instruction booklets had the same stimulus characteristics but the problems were varied by altering problem order and the attribute labels associated with each information configuration. For example, as Column 1 in **Table 1** shows, the Bayesian probabilities for the encounter with mystery person 5 in the pre-instruction booklet are identical to those for the encounter with mystery person two in the post-instruction booklet. Thus, task difficulty was precisely matched between pre- and post-instruction testing sessions.

Design

The stimulus characteristics shown in **Table 1** take the form of a 2 (Feature: present, absent) × 2 (Expected Error Direction: subadditive, superadditive) × 2 (Expected Error Magnitude: smaller, larger) within-subjects factorial design. The values of the first factor are shown in Column 2 of **Table 1**. The values of the second factor are encoded in column 7, where the values 0.44 and 0.80 indicate that subadditive judgments are expected if subjects commit the inverse fallacy and where the values 1.20 and 1.56 indicate that superadditive judgments are expected if subjects commit the inverse fallacy. The values 0.80 and 1.20 represent the smaller predicted errors, whereas the values 0.44 and 1.56 represent the larger predicted errors. Taking the pre-post manipulation into account, the experiment utilizes a 2 (Instruction) × 2 (Feature) × 2 (Expected Error Direction) × 2 (Expected Error Magnitude) within-subjects factorial design.

Results

Experience, as indexed by the level of course taken (i.e., 1 = junior, 2 = intermediate, and 3 = senior), was not significantly correlated with bias ($r = -0.07$, $p = 0.67$) or absolute bias (i.e., the degree of inaccuracy irrespective of whether it represents under- or over-estimation; $r = -0.15$, $p = 0.33$). Thus, experience is not statistically controlled in subsequent analyses.

Probability Judgment

To avoid redundancy in the presentation of the results, analyses are conducted on the additivity of probability judgments for Groups A and B. The statistical analyses accompanying these analyses are, of necessity, identical in inferential characteristics, such as significance levels and effect sizes, to those focusing instead on mean bias as a measure of inaccuracy, where bias is defined as the deviation between subjects' probability judgments and the estimates based on Bayes theorem. For instance, where $T' = 0.44$ or 1.56, a subject who invariably uses the inverse strategy would show a bias in his or her forecasts equal to $|0.56|$. Likewise, the subject would show an additivity violation, whereby T (i.e., the sum of his or her judgments for Groups A and B) would either exceed (when $T' = 1.56$) or fall short (when $T' = 0.44$) of unity by the same degree (i.e., 0.56).

Subjects' T-values were analyzed in a 2 (Instruction) × 2 (Feature) × 2 (Expected Error Direction) × 2 (Expected Error Magnitude) within-subjects factorial analysis of variance (ANOVA) model. There was a significant and large instruction effect showing that the additivity (and, by implication, mean agreement with Bayes theorem) of subjects' judgments improved from pre-instruction ($M = 0.91$, $SE = 0.028$) to post-instruction ($M = 0.99$, $SE = 0.008$) testing, $F_{(1,\ 42)} = 6.82$, $p = 0.012$, $\eta_p^2 = 0.14$. As the estimated marginal means show, prior to instruction, subjects' judgments, on average, were subadditive.

TABLE 1 | Summary of stimulus characteristics in judgment task.

Task no. (pre, post)	*D*	$P(D_+\|H_A)$	$P(D_+\|H_{\neg A})$	$P(D\|H_A)$	$P(D\|H_{\neg A})$	T'	$P(H_A\|D)$	$P(H_{\neg A}\|D)$
5, 2	Present	0.42	0.02	0.42	0.02	0.44	0.95	0.05
6, 1	Absent	0.58	0.98	0.42	0.02	0.44	0.95	0.05
8, 3	Absent	0.40	0.80	0.60	0.20	0.80	0.75	0.25
7, 4	Present	0.60	0.20	0.60	0.20	0.80	0.75	0.25
3, 8	Present	0.80	0.40	0.80	0.40	1.20	0.67	0.33
4, 7	Absent	0.20	0.60	0.80	0.40	1.20	0.67	0.33
1, 6	Present	0.98	0.58	0.98	0.58	1.56	0.63	0.37
2, 5	Absent	0.02	0.42	0.98	0.58	1.56	0.63	0.37

D_+, target has attribute; D, the result for the target (either has or doesn't have attribute); H_A, hypothesis that target belongs to Group A; $H_{\neg A}$, hypothesis that target belongs to Group B. $T' = P(D|H_A) + P(D|H_{\neg A})$.

TABLE 2 | Estimated mean *T*-values by instruction and expected error direction.

Expected	Instruction					
Error	Pre			Post		
Direction	*M*	LB	UB	*M*	LB	UB
Subadditive	0.83	0.75	0.90	0.95	0.91	0.99
Superadditive	0.99	0.91	1.07	1.02	0.99	1.05

LB and UB = 95% CI lower and upper bounds, respectively.

As predicted, the effect of instruction on additivity was moderated by the expected error direction, $F_{(1,\ 42)} = 10.13$, $p = 0.003$, $\eta_p^2 = 0.19$. **Table 2** shows the estimated marginal mean *T*-values with 95% confidence intervals (CI). As **Table 2** shows, instruction had a strong, beneficial effect on tasks in which subadditivity was predicted, $F_{(1,\ 42)} = 10.27$, $p = 0.003$, $\eta_p^2 = 0.20$. In that task subset, subadditivity was virtually eliminated post-instruction. In contrast, instruction had no effect when superadditivity was expected ($F < 1$). However, given that superadditivity was not found, the null effect of instruction in that context is to be expected. Rather, in that context, subjects' judgments, on average, were additive before and after instruction. No other effect in the full factorial model was significant at $p < 0.05$.

The former additivity analyses showed that subjects' judgments were subadditive, which implies that, on average, they underestimated the normative estimates. As **Table 1** shows, $P(D|H_A) > P(D|H_{\neg A})$ and, likewise, $P(H_A|D) > P(H_{\neg A}|D)$. Thus, one might expect that bias expressed in absolute terms would be more pronounced for judgments of $P(H_A|D)$ than judgments of $P(H_{\neg A}|D)$. To test this hypothesis, the absolute deviation between judged and normative probabilities were analyzed in a 2 (Instruction) × 2 [Judgment type: $P(H_A|D)$, $P(H_{\neg A}|D)$] within-subjects ANOVA. In fact, mean absolute bias was greater for judgments of $P(H_A|D)$ ($M = 0.144$, $SE = 0.013$) than judgments of $P(H_{\neg A}|D)$ ($M = 0.009$, $SE = 0.011$), $F_{(1,\ 42)} = 12.38$, $p = 0.001$, $\eta_p^2 = 0.23$. As **Figure 2** shows, judgment type also moderated the effect of instruction, such that there was a greater effect for judgments of $P(H_A|D)$ than judgments of $P(H_{\neg A}|D)$, $F_{(1,\ 42)} = n6.67$, $p = 0.013$, $\eta_p^2 = 0.14$. In other words, instruction had a greater effect on bias reduction (i.e., improving agreement with Bayes theorem) where bias was greater to begin with.

The preceding analyses give additive analysts the benefit of the doubt. However, it is possible that some of the expressed additivity captured in this experiment is spurious. Karvetski et al. (2013) found that probability judgments of binary complements were often additive because subjects assigned values of 0.5 to $P(A)$ and $P(\neg A)$. This pattern—known as the fifty-fifty blip (Fischhoff and Bruine de Bruin, 1999)—is likely to reflect the subjects' deep epistemic uncertainty regarding the task. Given that subjects asked to judge probabilities are seldom given a "don't know" option, they tend to express that message by responding on the midpoint of the probability scale. And when they are given a "don't know" option, fifty-fifty responses are greatly reduced (Mandel, 2005, Experiment 1b).

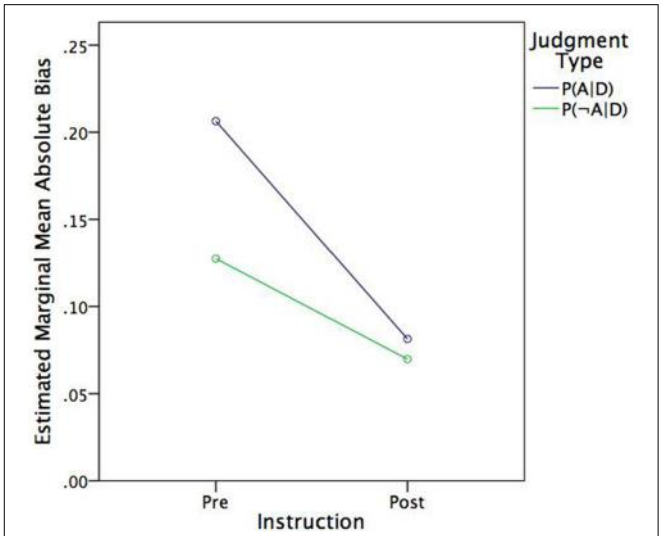

FIGURE 2 | Estimated marginal mean absolute bias by judgment type and instruction.

The pre- and post-instruction test data were scanned for fifty-fifty responders. Three subjects were spuriously additive in the pre-instruction test on at least five out of the eight problems. However, no subject showed this pattern in the post-instruction test. Thus, the prior results slightly underestimate the positive instruction effect by including the spurious cases of additive judgment in the pre-instruction test phase. Deletion of the three subjects, however, had no substantial effect on the results. The main effect of instruction on subjects' *T*-values was virtually unchanged, $F_{(1,\ 39)} = 6.89$, $p = 0.012$, $\eta_p^2 = 0.15$; and likewise for the instruction × direction interaction effect, $F_{(1,\ 39)} = 10.30$, $p = 0.003$, $\eta_p^2 = 0.21$. **Figure 3** shows the distribution of mean *T*-values before and after instruction with the three fifty-fifty responders excluded. It is evident that instruction was effective in improving the performance of the worst performers. In fact, the range post-instruction was less than one-third of its pre-instruction value (range = 0.26 vs. 0.88, respectively).

After removing the cases of spurious additivity, it is also of interest to compare the mean proportion of additive probability judgments before and after instruction. Instruction had a large effect on the mean proportion of additive judgments, which was greater after instruction ($M = 0.56$, $SD = 0.42$) than before instruction ($M = 0.75$, $SD = 0.31$), $t_{(39)} = 2.86$, $p = 0.007$, Cohen's $d = 0.91$. The proportion of subjects who were consistently additive across the eight problems in a test session was substantially greater after instruction (0.83) than before instruction (0.54)—a 54% increase in consistently additive responding by subjects.

Binary Choice

Although the tutorials used in this experiment did not mention choice, it was of interest to examine whether instruction may also have had a beneficial effect on the coherence of binary choices subjects made regarding the group to which the target

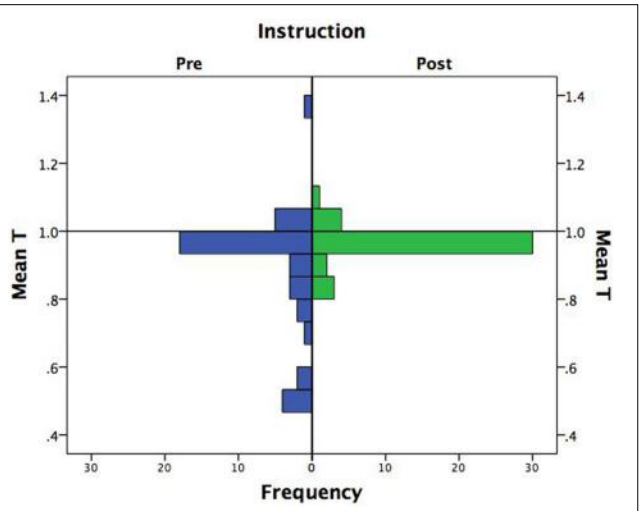

FIGURE 3 | Frequency distribution of additivity values (*T*) by instruction.

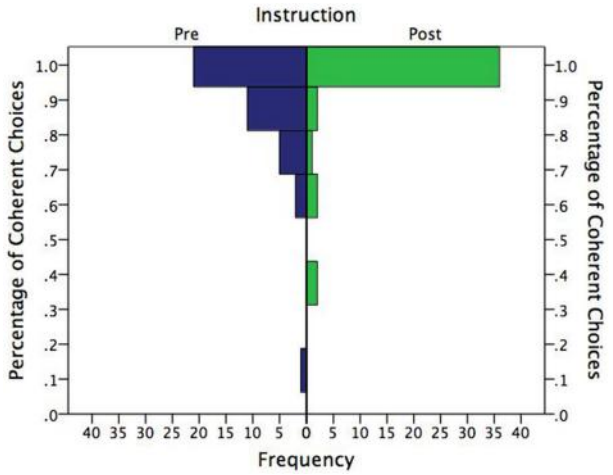

FIGURE 4 | Frequency distribution of percentage of coherent choices by instruction.

belonged. Coherent choices are defined as those in which the subject chooses the category as the target's group to which he or she assigned the higher probability. Conversely, if the subject chooses the group to which he or she assigned the lower probability, the choice is defined as incoherent.

Figure 4 shows the distribution of correct choices in percentage terms by instruction. Unsurprisingly, the distributions are highly skewed, with most subjects choosing coherently in all eight problems. What may be somewhat surprising, however, is that these distributions were not even more skewed. Clearly, the pre-instruction group showed considerable room for improvement, and improve with instruction they did. The proportion who chose coherently in all eight problems vs. those who made at least one incoherent choice was significantly greater after instruction (83%) than prior to instruction (53%), two-tailed binomial $p = 0.002$.

Discussion

The present research adds to the body of literature showing that Bayesian reasoning can be improved through relatively brief instruction in how to structure information using natural frequency representations (Sedlmeier, 1999; Sedlmeier and Gigerenzer, 2001; Kurzenhäuser and Hoffrage, 2002; Ruscio, 2003; McCloy et al., 2007; Sirota et al., 2015). In the present experiment, brief instruction in how to represent base-rate and diagnostic probabilities as natural-frequency-tree diagrams and how to then select the relevant subsets for calculation led to a large improvement in the additivity of intelligence analysts' posterior probability judgments of binary complements. As noted earlier, this effect also reflects the degree to which those probability judgments corresponded with those given by Bayes theorem.

Consistent with the majority of previous studies that have examined violations of the complementarity constraint (Macchi et al., 1999; Baratgin and Noveck, 2000; Windschitl et al., 2003, Experiment 4; Sloman et al., 2004; Mandel, 2005; Williams and Mandel, 2007; Mandel, 2008, Experiments 5 and 6), subjects' judgments were, on average, subadditive. Nevertheless, the results also show that most subjects were consistently additive in both pre- and post- instruction test phases, with a substantial rise in that proportion after instruction. Indeed, over four-fifths of subjects answered all eight problems additively after receiving instruction. What is also striking is that over half of them did so even before receiving instruction. It is likely that these proportions were as high as they were because the binary complements were elicited in immediate succession. Prior studies (Mandel, 2005; Karvetski et al., 2013) have found that spacing binary complements apart with unrelated items or tasks reduces the likelihood of additive responses. Thus, the proportions of consistently coherent subjects obtained in this research should be interpreted as having been elicited under near ideal conditions (short of prompting subjects to make their related judgments sum to unity; e.g., see Baratgin and Noveck, 2000). It would be of value to assess the effect of instruction on additivity when the binary complements are elicited in a spaced design.

The findings also showed that the degree of subadditivity manifested across pre-instruction problem sets was consistent with use of the inverse fallacy. That is, when the inverse (i.e., diagnostic) probabilities summed to less than unity ($T' < 1$), judgments were subadditive. In contrast, when the inverse probabilities summed to more than unity ($T' > 1$), the pre-instruction judgments were additive—and significantly less subadditive. Nevertheless, the results of this experiment do not confirm subjects' commission of the inverse fallacy as strongly as the findings obtained by Villejoubert and Mandel (2002) because, unlike their findings which showed superadditivity when $T' > 1$, the present findings revealed additive judgment under this condition. Simply put, exclusive reliance on the inverse fallacy in the present task would not have led to overall subadditivity.

An encouraging result was that instruction benefitted intelligence analysts' probability judgments where it was needed most. First, the effect of instruction was appropriately restricted to the subset of problems in which the inverse probabilities summed to less than unity. Under those conditions, instruction reduced additivity violations. However, for the $T' > 1$ task subset, where subjects' judgments were additive, instruction had no effect. This null simple effect is an important result because it shows that instruction did not merely make subjects' assigned probabilities larger across the board, as some other interventions appear to have done (e.g., Williams and Mandel, 2007). The assigned probabilities only became larger where they ought to have become larger. In other words, the benefit of instruction was appropriately targeted. Second, the effect of instruction on reducing mean absolute bias was greatest for the set of judgments that yielded the greatest absolute bias in the pre-instruction test (i.e., $P(H_A|D)$).

The benefit of instruction, as noted earlier, was also targeted in the sense that those who performed relatively poorly on the pre-instruction test, showed clear signs of improvement, as indicated by the large reduction in the range of performance post-instruction as compared to pre-instruction. This was evident in terms of both violation of the complementarity constraint and coherence of binary choices. Moreover, the few analysts who provided fifty-fifty responses prior to instruction no longer did so after instruction. These results are promising because they indicate that large improvements in probability judgment, information integration, and belief revision can be made by those who need improvement the most. Of course, the present research cannot speak to the long-term effect of instruction because the post-instruction test was administered immediately after training. However, as noted earlier, a number of studies have shown long-term beneficial effects on Bayesian judgment of instruction that has relied on the use of natural frequency representations of evidence (e.g., Sedlmeier and Gigerenzer, 2001). It would nevertheless be useful to confirm that there is a long-term benefit to judgmental coherence and also that such benefits can be derived from experts who are tasked with making judgments under conditions of uncertainty (such as intelligence analysts).

Likewise, given the encouraging results of this and other research on the use of instruction to improve aspects of Bayesian judgment, it would be of value to explore how such instruction might be further optimized by incorporating other effective learning techniques (for overviews, see Dunlosky et al., 2013; Kober, 2015). For instance, most studies of instruction effects on Bayesian reasoning, including the present research, have used a massed training and practice session. However, much experimental evidence indicates that students learn more effectively when they are given opportunities for distributed practice with large time lags between sessions (Cepeda et al., 2006; Delaney et al., 2010). While the majority of studies have demonstrated the benefits of distributed practice using factual materials that require mainly recall ability, Kapler et al. (2014) have shown that distributed practice in a simulated undergraduate classroom improves learning of higher-level reasoning that requires both recall and manipulation of information, much as Bayesian reasoning requires.

Finally, it is worth noting that the present research yielded not only a large statistical effect but also a practical effect given that the instructional method developed and tested in this research has since been adopted in some intelligence courses in Canada. Of course, it remains unclear to what extent such training will ultimately affect the quality of intelligence analysis and whether, in fact, Bayesianism is an appropriate model for belief revision in that domain (for an insightful discussion, see Zlotnick, 1972). Given that most assessments are communicated with verbal probability phrases and few assessments are based on evidence for which uncertainties are quantified, the application of aspects of Bayesianism such as Bayes theorem are currently of limited value. Nevertheless, even verbal probabilities should respect coherence principles such as additivity. It may be more difficult to verify whether "very likely that *A* will happen" and "slim chance that *A* won't happen" add up to unity, and such verification would be less direct because it would require personally translating the phrases into numbers. However, even without translation attempts, one could be reasonably confident that "almost certain that it's *A*" and "fifty-fifty that it's not *A*" are superadditive. Moreover, judgment accuracy is substantially improved by giving subjects in an opinion pool weight proportional to their adherence to the additivity principle (Karvetski et al., 2013). Forecast accuracy has also been improved by probability training that took the form of directives and rules of thumb aimed at avoiding common pitfalls, such as assigning probabilities of fifty-fifty to binary complements when forecasters are deeply unsure (Mellers et al., 2014). The instructional method developed in this research could potentially be used on its own or in combination with directive-based probability training to improve the quality of forecasting in the intelligence community and in other expert domains requiring probability judgment.

Acknowledgments

I thank Ron Wulf for facilitating the research at the Canadian Forces School for Military Intelligence and for developing the intelligence version of the tutorial. I also thank Natalia Derbentseva and Lianne McLellan for assistance in conducting this research. Funding for this research was provided by DRDC Applied Research Program Project 15dm "Understanding and Augmenting Human Analytic Capabilities" and by the DRDC Joint Intelligence Collection and Capabilities Project.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2015.00387/abstract

References

Adams, B. A., Thomson, M., Derbentseva, N., and Mandel, D. R. (2012). *Capability Challenges in the Human Domain for Intelligence Analysis: Report on Community-Wide Discussions with Canadian Intelligence Professionals [Contractor Report CR-2011-182]*. Toronto, ON: Defence Research and Development Canada.

Ayal, S., and Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgm. Decis. Mak*. 9, 226–242.

Baratgin, J., and Noveck, I. (2000). Not only base-rates are neglected in the Lawyer-Engineer problem: an investigation of reasoners' underutilization of complementarity. *Mem. Cogn.* 28, 79–91. doi: 10.3758/BF03211578

Baratgin, J., Over, D. E., and Politzer, G. (2014). New psychological paradigm for conditionals and General de Finetti Tables. *Mind Lang.* 29, 73–84. doi: 10.1111/mila.12042

Baratgin, J., and Politzer, G. (2006). Is the mind Bayesian? The case for agnosticism. *Mind Soc*. 5, 1–38. doi: 10.1007/s11299-006-0007-1

Baratgin, J., and Politzer, G. (2010). Updating: a psychologically basic situation of probability revision. *Think. Reason.* 16, 253–287. doi: 10.1080/13546783.2010.519564

Baratgin, J. (2009). Updating our beliefs about inconsistency: the Monty-Hall case. *Math. Soc. Sci*. 57, 67–95. doi: 10.1016/j.mathsocsci.2008.08.006

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–297. doi: 10.1017/S0140525X07001653

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

Brase, G. L. (2008). Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol.* 23, 369–381. doi: 10.1002/acp.1460

Brase, G. L. (2014). The power of representation and interpretation: doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *J. Cogn. Psychol.* 26, 81–97. doi: 10.1080/20445911.2013.861840

Casscells, W., Schoenberger, A., and Graboys, T. (1978). Interpretation by physicians of clinical laboratory results. *N. Engl. J. Med.* 299, 999–1001. doi: 10.1056/NEJM197811022991808

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., and Rohrer, D. (2006). Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychol. Bull.* 132, 354–380. doi: 10.1037/0033-2909.132.3.354

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

Cozic, M. (2011). Imaging and sleeping beauty: the case for double-halfers. *Int. J. Approx. Reason*. 52, 147–153. doi: 10.1016/j.ijar.2009.06.010

Delaney, P. F., Verkoeijen, P. P. J. L., and Spirgel, A. (2010). Spacing and the testing effects: a deeply critical, lengthy, and at times discursive review of the literature. *Psychol. Learn. Motiv.* 53, 63–147. doi: 10.1016/S0079-7421(10)53003-2

Douven, I., and Schupbach, J. N. (2015). Probabilistic alternatives to Bayesianism: the case of explanationism. *Front. Psychol.* 6:459. doi: 10.3389/fpsyg.2015.00459

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., and Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol. Sci. Public Interest* 14, 4–58. doi: 10.1177/1529100612453266

Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine: problems and opportunities," in *Judgment under Uncertainty: Heuristics and Biases*, eds D. Kahneman, P. Slovic and A. Tversky (New York, NY: Cambridge University Press), 249–267.

Evans, J. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Think. Reason.* 18, 5–31. doi: 10.1080/13546783.2011.637674

Fischhoff, B., and Bruine de Bruin, W. (1999). Fifty-fifty=50%? *J.* Behav. *Decis. Mak*. 12, 149–163. doi: 10.1002/(SICI)1099-0771(199906)12:2<149::AID-BDM314>3.0.CO;2-J

Friedman, J. A., and Zeckhauser, R. (2012). Assessing uncertainty in intelligence. *Intell. Natl. Secur.* 27, 824–847. doi: 10.1080/02684527.2012.708275

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

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev*. 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., and Hoffrage, U. (2007). The role of representation in Bayesian reasoning: correcting common misconceptions. *Behav. Brain Sci.* 30, 264–267. doi: 10.1017/S0140525X07001756

Gigerenzer, G., Hoffrage, U., and Ebert, A. (1998). AIDS counselling for low-risk clients. *AIDS Care* 10, 197–211. doi: 10.1080/09540129850124451

Hamm, R. (1993). Explanations for common responses to the Blue/Green cab probabilistic inference word problem. *Psychol. Rep.* 72, 219–242. doi: 10.2466/pr0.1993.72.1.219

Hoffrage, U., and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Acad. Med.* 73, 538–540. doi: 10.1097/00001888-199805000-00024

Juslin, P., Winman, A., and Olsson, H. (2003). Calibration, additivity, and source independence of probability judgments in general knowledge and sensory discrimination tasks. *Organ. Behav. Hum. Decis. Process.* 92, 34–51. doi: 10.1016/S0749-5978(03)00063-3

Juslin, P. (2015). Controlled information integration and Bayesian inference. *Front. Psychol.* 6:70. doi: 10.3389/fpsyg.2015.00070

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol*. 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychol. Rev*. 80, 237–251. doi: 10.1037/h0034747

Kapler, I. V., Weston, T., and Wiseheart, M. (2014). Spacing in a simulated undergraduate classroom: long-term benefits for factual and higher-level learning. *Learn. Instr*. 36, 38–45. doi: 10.1016/j.learninstruc.2014.11.001

Karvetski, C. W., Olson, K. C., Mandel, D. R., and Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decis. Anal.* 10, 305–326. doi: 10.1287/deca.2013.0279

Kent, S. (1964). *Words of Estimative Probability*. Available online at: https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/sherman-kent-and-the-board-of-national-estimates-collected-essays/6words.html

Kleiter, G. (1994). "Natural sampling: rationality without base rates," in *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, eds G. H. Fischer and D. Laming (New York, NY: Springer-Verlag), 375–388.

Kober, N. (2015). *Reaching Students: What Research says about Effective Instruction in Undergraduate Science and Engineering. Board on Science Education, Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

Koehler, J. J. (1996). The base rate fallacy reconsidered: descriptive, normative and methodological challenges. *Behav. Brain Sci.* 19, 1–17. doi: 10.1017/S0140525X00041157

Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability*. Oxford, UK: Chelsea Publishing Co.

Kurzenhäuser, S., and Hoffrage, U. (2002). Teaching Bayesian reasoning: an evaluation of a classroom tutorial for medical students. *Med. Teach*. 24, 516–521. doi: 10.1080/0142159021000012540

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philos. Rev*. 85, 297–315. doi: 10.2307/2184045

Lyon, D., and Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychol.* 40, 287–298. doi: 10.1016/0001-6918(76)90032-9

Macchi, L., Osherson, D., and Krantz, D. H. (1999). A note on superadditive probability judgment. *Psychol. Rev.* 106, 210–214. doi: 10.1037/0033-295X.106.1.210

Macchi, L. (1995). Pragmatic aspects of the base rate fallacy. *Q. J. Exp. Psychol.* 48A, 188–207. doi: 10.1080/14640749508401384

Mandel, D. R., and Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proc. Nat. Acad. Sci. U.S.A.* 111, 10984–10989. doi: 10.1073/pnas.1406138111

Mandel, D. R. (2005). Are risk assessments of a terrorist attack coherent? *J. Exp. Psychol. Appl*. 11, 277–288. doi: 10.1037/1076-898X.11.4.277

Mandel, D. R. (2007). Nested-sets theory, full stop: explaining performance on Bayesian inference tasks without dual-systems assumptions. *Behav. Brain Sci.* 30, 275–276. doi: 10.1017/S0140525X07001835

Mandel, D. R. (2008). Violations of coherence in subjective probability: a representational and assessment processes account. *Cognition* 106, 130–156. doi: 10.1016/j.cognition.2007.01.001

Mandel, D. R. (2014a). The psychology of Bayesian reasoning. *Front. Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

Mandel, D. R. (2014b). Visual representation of rational belief revision: another look at the Sleeping Beauty problem. *Front. Psychol*. 5:1232. doi: 10.3389/fpsyg.2014.01232

McCloy, R., Beaman, C. P., Morgan, B., and Speed, R. (2007). Training conditional and cumulative risk judgements: the role of frequencies, problem-structure and Einstellung. *Appl. Cogn. Psychol.* 21, 325–344. doi: 10.1002/acp.1273

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., et al. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychol. Sci.* 25, 1106–1115. doi: 10.1177/0956797614524255

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic approach to Human Reasoning*. Oxford, UK: Oxford University Press.

Pool, R. (2010). *Field Evaluation in the Intelligence and Counterintelligence Context: Workshop Summary*. Washington, DC: The National Academies Press.

Rottenstreich, Y., and Tversky, A. (1997). Unpacking, repacking, and anchoring: advances in support theory. *Psychol. Rev*. 104, 406–415. doi: 10.1037/0033-295X.104.2.406

Ruscio, J. (2003). Comparing Bayes's theorem to frequency-based approaches to teaching Bayesian reasoning. *Teach. Psychol.* 30, 325–328.

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen*. 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Sedlmeier, P. (1999). *Improving Statistical Reasoning: Theoretical Models and Practical Implications*. Mahwah, NJ: Erlbaum.

Sirota, M., Kostovičová, L., and Vallée-Tourangeau, F. (2015). How to train your Bayesian: a problem-representation transfer rather than a format-representation shift explains training effects. *Q. J. Exp. Psychol.* 68, 1–9. doi: 10.1080/17470218.2014.972420

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., and Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *J. Exp. Psychol. Learn. Mem. Cognit.* 30, 573–582. doi: 10.1037/0278-7393.30.3.573

Sloman, S. A., Over, D. E., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Process*. 91, 296–309. doi: 10.1016/S0749-5978(03)00021-9

Thagard, P. (1989). Explanatory coherence. *Behav. Brain Sci.* 12, 435–502. doi: 10.1017/S0140525X00057046

Tversky, A., and Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychol. Rev*. 101, 547–567. doi: 10.1037/0033-295X.101.4.547

Villejoubert, G., and Mandel, D. R. (2002). The inverse fallacy: an account of deviations from Bayes's theorem and the additivity principle. *Mem. Cogn*. 30, 171–178. doi: 10.3758/BF03195278

Walliser, B., and Zwirn, D. (2002). Can Bayes' rule be justified by cognitive rationality principles? *Theory Decis.* 53, 95–135. doi: 10.1023/A:102122710674

Wallsten, T. S., Budescu, D. V., and Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Manage. Sci*. 39, 176 –190. doi: 10.1287/mnsc.39.2.176

Williams, J. J., and Mandel, D. R. (2007). "Do evaluation frames improve the quality of conditional probability judgment?," in *Proceedings of the 29th Annual Meeting of the Cognitive Science Society,* eds D. S. McNamara and J. G. Trafton (Mahwah, NJ: Erlbaum), 1653–1658.

Windschitl, P. D., Kruger, J., and Simms, E. N. (2003). The influence of egocentrism and focalism on people's optimism in competitions: when what affects us equally affects me more. *J. Pers. Soc. Psychol.* 85, 398–408. doi: 10.1037/0022-3514.85.3.389

Wolfe, C. R. (1995). Information seeking on Bayesian conditional probability problems: a fuzzy-trace theory account. *J. Behav. Decis. Mak*. 8, 85–108. doi: 10.1002/bdm.3960080203

Zlotnick, J. (1972). Bayes theorem for intelligence analysis. *Stud. Intell.* 16, 43–52.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 26 August 2015
doi: 10.3389/fpsyg.2015.01186

Effects of visualizing statistical information – an empirical study on tree diagrams and 2 × 2 tables

Karin Binder, Stefan Krauss and Georg Bruckmaier*

Mathematics Education, Faculty of Mathematics, University of Regensburg, Regensburg, Germany

In their research articles, scholars often use 2 × 2 tables or tree diagrams including natural frequencies in order to illustrate Bayesian reasoning situations to their peers. Interestingly, the effect of these visualizations on participants' performance has not been tested empirically so far (apart from explicit training studies). In the present article, we report on an empirical study (3 × 2 × 2 design) in which we systematically vary visualization (no visualization vs. 2 × 2 table vs. tree diagram) and information format (probabilities vs. natural frequencies) for two contexts (medical vs. economical context; not a factor of interest). Each of $N = 259$ participants (students of age 16–18) had to solve two typical Bayesian reasoning tasks ("mammography problem" and "economics problem"). The hypothesis is that 2 × 2 tables and tree diagrams – especially when natural frequencies are included – can foster insight into the notoriously difficult structure of Bayesian reasoning situations. In contrast to many other visualizations (e.g., icon arrays, Euler diagrams), 2 × 2 tables and tree diagrams have the advantage that they can be constructed easily. The implications of our findings for teaching Bayesian reasoning will be discussed.

Keywords: Bayesian reasoning, 2 × 2 table, natural sampling tree, natural frequencies, visual representation

OPEN ACCESS

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
Miroslav Sirota, Kingston University London, UK
Artur Domurat, University of Warsaw, Poland

****Correspondence:***
Karin Binder, Mathematics Education, Faculty of Mathematics, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany
karin.binder@ur.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *30 March 2015*
Accepted: *27 July 2015*
Published: *26 August 2015*

Citation:
Binder K, Krauss S and Bruckmaier G (2015) Effects of visualizing statistical information – an empirical study on tree diagrams and 2 × 2 tables. Front. Psychol. 6:1186. doi: 10.3389/fpsyg.2015.01186

Introduction

Bayes' formula is vitally important in many areas, such as in medicine or law. Unfortunately, both laymen and professionals have trouble understanding and combining statistical information effectively. The resulting misjudgments can have severe consequences, for example when juries must convict or acquit defendants based on probabilistic evidence in legal trials (Hoffrage et al., 2000; Krauss and Bruckmaier, 2014), or when physicians have to understand and to communicate what a positive test result really means, for example in a HIV or cancer test (Ellis et al., 2014). Consider, for instance, the classic mammography problem (adapted from Eddy, 1982; see also Gigerenzer and Hoffrage, 1995; Siegrist and Keller, 2011; Micallef et al., 2012; Garcia-Retamero and Hoffrage, 2013).

Mammography Problem (Probability Format):

The probability of breast cancer is 1% for a woman who participates in routine screening. If a woman who participates in routine screening has breast cancer, the probability is 80% that she will have a positive test result. If a woman who participates in routine screening does not have breast cancer, the probability is 9.6% that she will have a positive test result. What is the probability that a woman who participates in routine screening and receives a positive test result has breast cancer?
Answer: _____ %

According to Bayes' theorem, the resulting posterior probability P(B| M+) is:

$$P(B|M+) = \frac{P(M+|B) \cdot P(B)}{P(M+|B) \cdot P(B) + P(M+|\neg B) \cdot P(\neg B)} = \frac{80\% \cdot 1\%}{80\% \cdot 1\% + 9.6\% \cdot 99\%} \approx 7.8\%$$

The correct result 7.8% is much lower than most people, including physicians, would expect (Eddy, 1982). Several studies show that medical doctors (Hoffrage and Gigerenzer, 1998; Garcia-Retamero and Hoffrage, 2013), patients (Garcia-Retamero and Hoffrage, 2013), legal professionals (Hoffrage et al., 2000), and students (Ellis et al., 2014) have difficulties with similar tasks. In order to help people to understand the situation, Gigerenzer and Hoffrage (1995) replaced the probabilities in Eddy's task by natural frequencies.

Mammography Problem (Natural Frequency Format):

> 100 out of 10,000 women who participate in routine screening have breast cancer. Out of 100 women who participate in routine screening and have breast cancer, 80 will have a positive result. Out of 9,900 women who participate in routine screening and have no breast cancer, 950 will also have a positive result. How many of the women who participate in routine screening and receive a positive test result have breast cancer?
> Answer: ____ out of ____

The percentage of correct responses increased from about 10–20% to about 50% in 15 different Bayesian reasoning tasks, including the mammography problem (Gigerenzer and Hoffrage, 1995). While the facilitating effect of natural frequencies is accepted by now, scholars differ in explaining this effect. Gigerenzer and Hoffrage (1995), for instance, argue that the human mind is evolutionarily adapted to the information format of natural frequencies ("ecological rationality") that result from a natural sampling process (Kleiter, 1994). Other theorists, however, claim that essentially the partitive information structure is responsible for the facilitating effect ("nested sets hypothesis"; e.g., Girotto and Gonzalez, 2001; Sloman et al., 2003; Barbey and Sloman, 2007). Some scholars suggest that two different cognitive systems ("dual process theory"; Sloman, 1996; Kahneman and Frederick, 2005; Barbey and Sloman, 2007) may be responsible for inferences with respect to the different information formats. While probability format triggers intuitive thinking according to system 1 ("associative system"; see also Sloman, 1996), which may lead to base rate neglect, natural frequency format triggers deliberate reasoning according to system 2 ("rule based system"). Advocates of the dual process theory often support the nested sets hypothesis (e. g., Barbey and Sloman, 2007). For a discussion of the concept of natural frequencies see Gigerenzer and Hoffrage (1999), Lewis and Keren (1999), Mellers and McGraw (1999), Girotto and Gonzalez (2001, 2002), Hoffrage et al. (2002), Barbey and Sloman (2007), or Sirota et al. (2015a).

In fact, there are recommendations that natural frequencies should become part of the training for all medical students (Gigerenzer, 2013) and, moreover, should be part of elementary school curricula (Gigerenzer, 2014). Although the effect of numerical format (probabilities vs. natural frequencies) is quite substantial, it has to be noted that there is still potential for improvement ("only" approximately 50% correct solutions).

Another idea to improve insight into Bayesian reasoning situations is the additional representation of visual aids such as *Euler diagrams*, *icon arrays*, *frequency grids*, *unit squares*, *roulette wheel diagrams*, and *tree diagrams* (see **Figure 1**). According to the nested sets hypothesis, most of these visual aids represent the set-subset relation of the information. For an overview of possible visualizations see Paling (2003) or Spiegelhalter et al. (2011). **Figure 1** shows some visual aids which have been tested empirically so far.

Sloman et al. (2003), Brase (2008), Micallef et al. (2012), and Sirota et al. (2014b) investigated to what extent the presentation of *Euler diagrams* can boost performance in Bayesian reasoning tasks. They obtained different findings regarding the effectiveness of Euler diagrams, a result which potentially is affiliated to the various types of participants in their studies. *Icon arrays* (also called *pictographs*) are matrices of small figures that represent the given information. Within an array, some of the icons are shaped in a special form or are colored in order to show the number of figures that fulfill a special feature. Brase (2008, 2014) and Zikmund-Fisher et al. (2014) recommend risk communication via icon arrays since their studies showed a positive influence of this visual aid (for a discussion of the concept of "iconicity" see, e.g., Sirota et al., 2014b). *Frequency grids* are close to icon arrays showing the overall number of persons in a large grid where particular subsets of persons are marked characteristically. Garcia-Retamero and Hoffrage (2013) found that both doctors' and patients' performance increased when frequency grids are provided (see also Garcia-Retamero et al., 2015). *Unit squares* (Bea, 1995; Sturm and Eichler, 2014) also mirror the statistical information geometrically and represent the different sets of the task. Bea (1995) recommends the visualization of information via a unit square since his research reveals substantial improvement in performance. *Roulette wheel diagrams* (Brase, 2014) summarize the information presented by two circles (inner and outer circle) which represent different subsets of the problem. However, the additional representation of a roulette wheel diagram causes only a very small or even no improvement in performance compared to versions without any visual aid (Brase, 2014). Friederichs et al. (2014) investigated *tree diagrams* without numerical values (except an imaginary sample size). In their studies, performance in probability versions with tree diagrams was similar to the performance in natural frequency versions without visualization.

Note that one can differentiate between two types of studies in general: On the one hand there are training studies where participants are explicitly instructed in how to create visual aids on their own, and consequently, how to combine the given numbers to arrive at the solution. The effect of this "teaching" then is investigated by presenting additional problems without

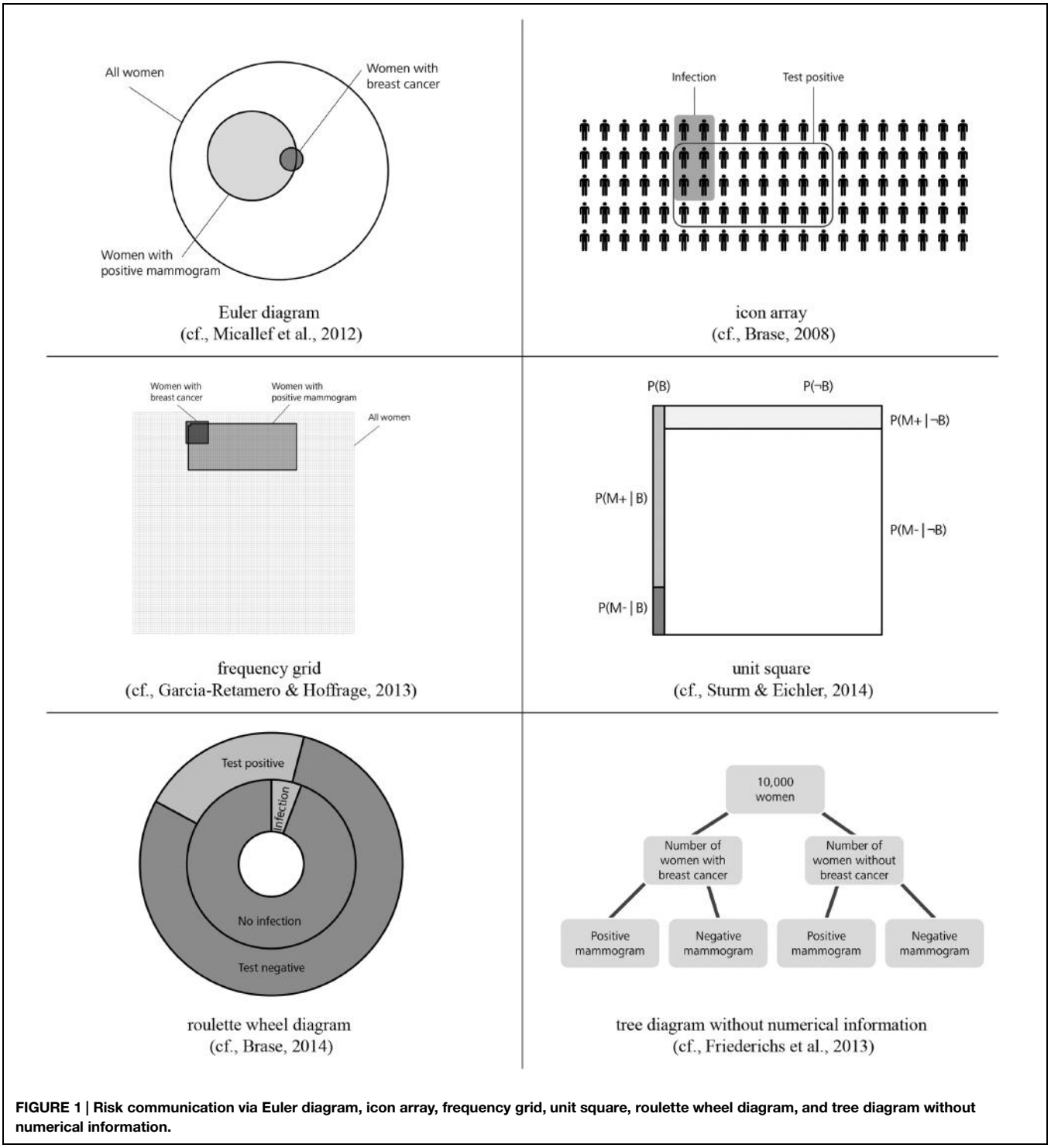

FIGURE 1 | Risk communication via Euler diagram, icon array, frequency grid, unit square, roulette wheel diagram, and tree diagram without numerical information.

visualizations (e.g., Sedlmeier and Gigerenzer, 2001; Ruscio, 2003; Sirota et al., 2015b). On the other hand there are studies – as in our study – where word problems are accompanied by visualizations (e.g., Brase, 2008; Garcia-Retamero and Hoffrage, 2013). Note that in the latter studies, it is *not* taught how to construct visualizations for fostering insight, and therefore, there is no prior instruction as to how the given numbers should be applied to infer the solution. The visualizations in this case rather illustrate the information of the given problem in parallel.

Interestingly, the beneficial effect of 2 × 2 tables and tree diagrams presently was investigated only in the context of training studies (e.g., Sedlmeier and Gigerenzer, 2001). This is astonishing since scholars commonly use tree diagrams (Kleiter,

1994; Gigerenzer and Hoffrage, 1995; Mandel, 2014; Navarrete et al., 2014) and 2 × 2 tables (Goodie and Fantino, 1996; Dougherty et al., 1999; Fiedler et al., 2000) containing numerical values in their research papers to represent Bayesian reasoning situations to their colleagues.

In the present paper we investigate how performance in Bayesian reasoning tasks can additionally be enhanced by providing 2 × 2 tables and tree diagrams containing numerical values. Since 2 × 2 tables and tree diagrams both can be equipped with natural frequencies or with probabilities we decided to test all four possible visualizations (compare **Figure 2**). Our hypotheses were:

- Hypothesis 1: Problems in which information is presented in natural frequencies are easier to solve than problems containing probabilities. This holds true when problems without visualization are compared (replication of previous studies) and when problems with visualizations are compared.
- Hypothesis 2: The additional presentation of visualizations of the numerical values (2 × 2 tables and tree diagrams) facilitates understanding. This holds for natural frequency and for probability versions as well.

We had no hypothesis as to which of both kinds of visualization is more beneficial. Furthermore we had no hypothesis on the effect of the problem context (we had chosen two problem contexts for mutual validation of our results; see **Table 1**).

Experimental Study

Design

In a paper-and-pencil questionnaire participants were presented with two Bayesian reasoning tasks, the mammography problem and a short version of the economics problem (Ajzen, 1977; for problem formulations see **Table 2**). The design of the study includes two factors of interest (visualization and format of information) and one factor which was not of interest (context), resulting in a 3 × 2 × 2 design:

- *Visualization:* no visualization vs. 2 × 2 table vs. tree diagram.
- *Format of statistical information:* probabilities vs. natural frequencies.
- *Context:* mammography problem vs. economics problem (not a factor of interest).

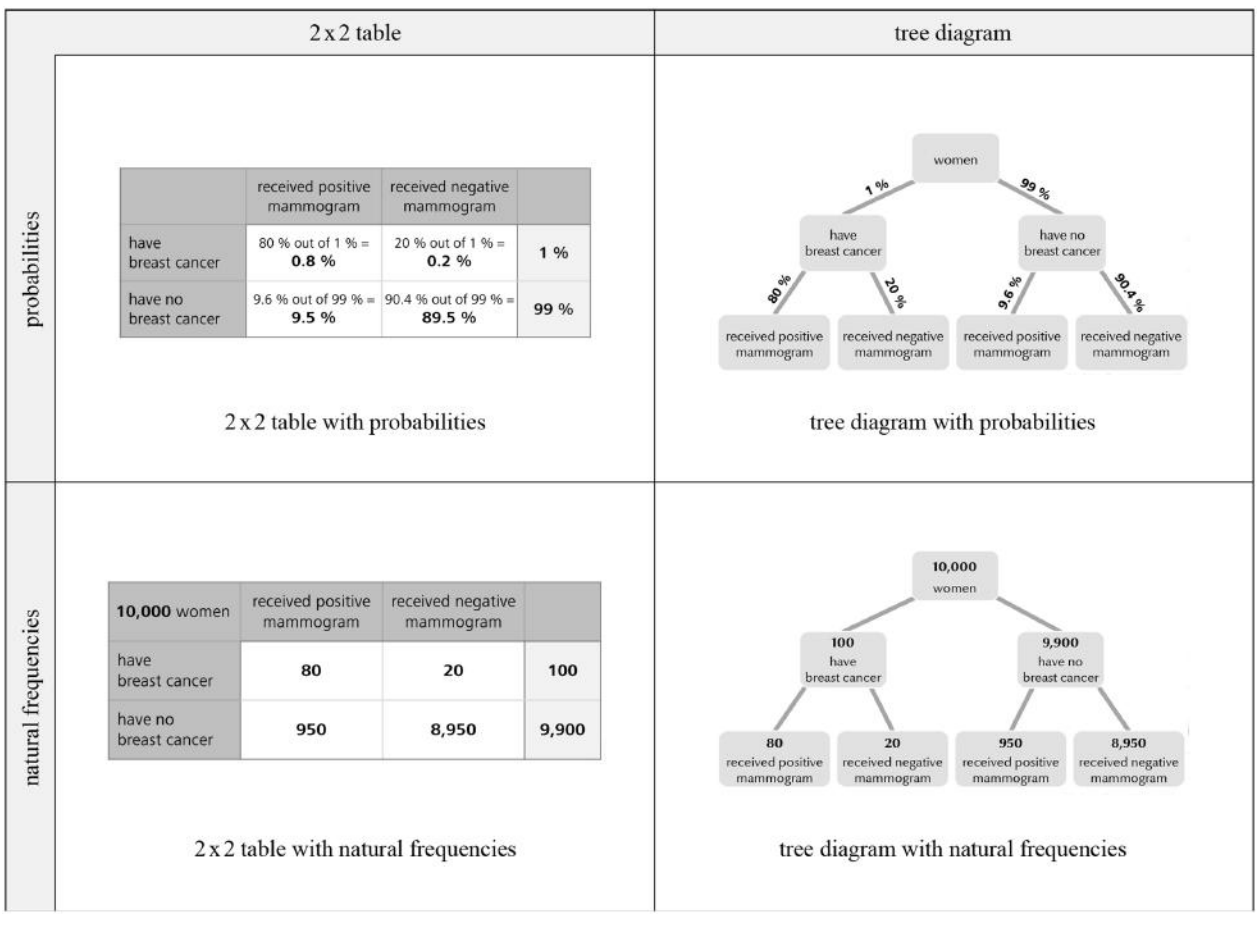

FIGURE 2 | Four resulting visualizations of the respective information format (mammography problem).

TABLE 1 | Design of the 12 tested problem versions.

		Context	
		Mammography problem	**Economics problem**
Format	**Probabilities**	• No visualization • 2 × 2 table • Tree diagram	• No visualization • 2 × 2 table • Tree diagram
	Natural frequencies	• No visualization • 2 × 2 table • Tree diagram	• No visualization • 2 × 2 table • Tree diagram

Each participant received one of the two problem contexts with probabilities and the other problem with natural frequencies. Thereby the order of context and information format was varied systematically. Furthermore, if in one of the two problems, for instance, a 2 × 2 table was added, in the other problem either no visualization or a tree diagram was presented. There were no time constraints for completing the questionnaire (participants required about 20 min for both tasks). In **Table 1** the design, resulting in 12 tested versions, is illustrated, whereas in **Table 2** the corresponding problem formulations are denoted.

The key factor under investigation in the present article is the effect of visualization. Note that in contrast to most visual aids tested so far (**Figure 1**) our visualizations explicitly contain numerical information. It is generally possible to equip both 2 × 2 tables and tree diagrams with natural frequencies or with probabilities, respectively (**Figure 2**). The construction rationale for the visualizations was to provide statistical information that is also reported in the typical problem formulations. However, to "complete" the tree diagrams some information must be added that is not mentioned in the problem formulation (the information "20%" and "90.4%" in the probability tree or "20" and "8,950" in the frequency tree, respectively). In order to mirror these numerical values in the 2 × 2 table containing natural frequencies, one (of two possible) marginal distribution has to be depicted (**Figure 2**). Most problematic is the construction of the 2 × 2 table with probabilities. Such 2 × 2 tables usually contain conjoint probabilities, whereas Bayesian reasoning tasks contain conditional probabilities. The underlying relationship between both kinds of probabilities is included in the cells of the 2 × 2 tables (probabilities). It has to be noted that the 2 × 2 table (with conjoint probabilities), the 2 × 2 table (with natural frequencies) and the tree diagram (with probabilities) are part of the German school curriculum, whereas the tree diagram with natural frequencies ("natural frequency tree")is not.

TABLE 2 | Problem formulations.

	Mammography problem		**Economics problem**	
	Probability version	**Natural frequency version**	**Probability version**	**Natural frequency version**
Cover story	Imagine you are a reporter for a women's magazine and you want to write an article about breast cancer. As a part of your research, you focuses on mammography as an indicator of breast cancer. You are especially interested in the question of what it means, when a woman has a positive result (which indicates breast cancer) in such a medical test. A physician explains the situation with the following information:		Imagine you are interested in the question, if career-oriented students are more likely to attend an economics course. Therefore the school psychological service evaluates the correlations of personality characteristics and choice of courses for you. The following information is available:	
Version	The probability of breast cancer is 1% for a woman who participates in routine screening. If a woman who participates in routine screening has breast cancer, the probability is 80% that she will have a positive test result. If a woman who participates in routine screening does not have breast cancer, the probability is 9.6% that she will have a positive test result.	100 out of 10,000 women who participate in routine screening have breast cancer. Out of 100 women who participate in routine screening and have breast cancer, 80 will have a positive result. Out of 9,900 women who participate in routine screening and have no breast cancer, 950 will also have a positive result.	The probability that a student attends the economics course is 32.5%. If a student attends the economics course, the probability that he is career oriented is 64%. If a student does not attend the economics course, the probability that he is still career-oriented is 60%.	325 out of 1,000 students attend the economics course. Out of 325 students who attend the economics course, 208 are career-oriented. Out of 675 students who not attend the economics course, 405 are still career-oriented.
Visual aid	• No visualization, or • 2 × 2 table (prob.), or • Tree diagram (prob.)	• No visualization, or • 2 × 2 table (nat. freq.), or • Tree diagram (nat. freq.)	• No visualization, or • 2 × 2 table (prob.), or • Tree diagram (prob.)	• No visualization, or • 2 × 2 table (nat. freq.), or • Tree diagram (nat. freq.)
Question	What is the probability that a woman who participates in routine screening and receives a positive test result has breast cancer? Answer: _____ %	How many of the women who participate in routine screening and receive a positive test result have breast cancer? Answer: ____ out of ____	What is the probability that a student attends the economics course if he is career-oriented? Answer: _____ %	How many of the students who are career-oriented attend the economics course? Answer: ____ out of ____

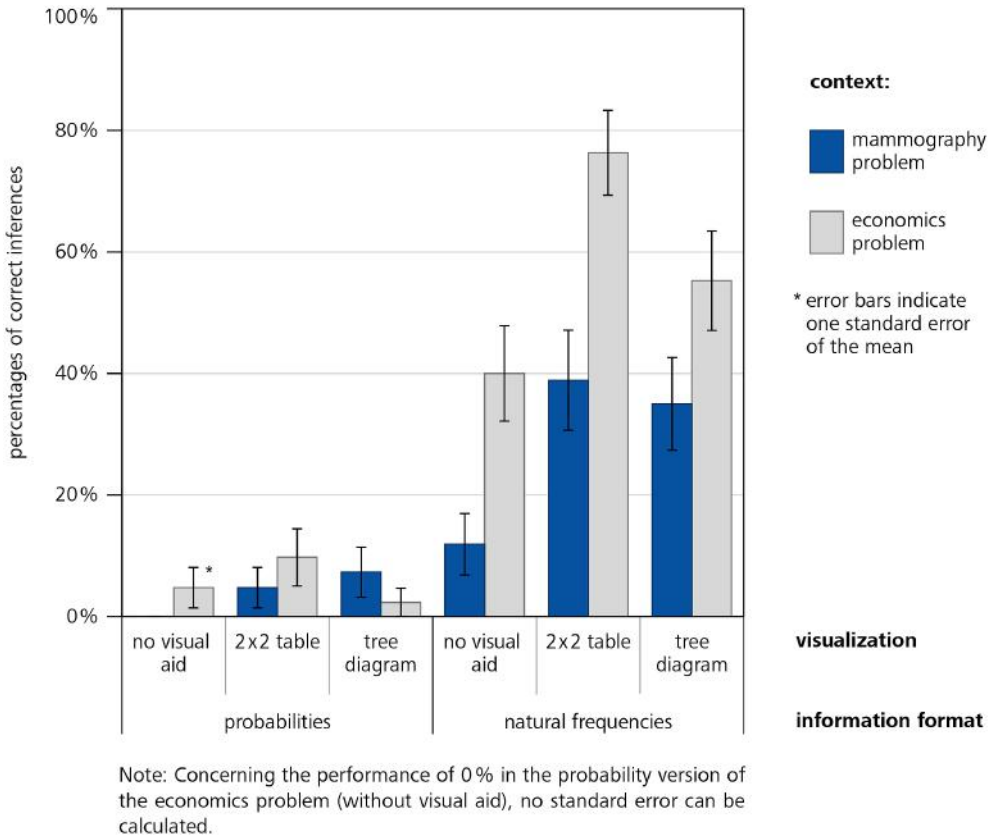

FIGURE 3 | Participants performance (error bars indicate the SE).

TABLE 3 | Results of binary logistic regression; independent variables: visualization and information format; dependent variable: correctness of solution.

	Dependent variable: correctness of solution			
	Mammography problem		Economics problem	
	Model 1	Model 2	Model 1	Model 2
Independent variable	EXP(B)	EXP(B)	EXP(B)	EXP(B)
Format of information	9.40***	10.44***	22.44***	24.73***
Visualization		4.99**		2.53*
R^2	0.19	0.27	0.41	0.44

EXP(B): Odds ratio (indicates how many times the odds of solving the task is higher when the independent variable is 1, as compared to the independent variable of 0); R^2: Goodness of fit (according to Nagelkerke).
**significant at p = 0.05; **significant at p = 0.01; ***significant at p = 0.001.*

Instrument

Each participant was presented two successive tasks which varied in terms of (1) visualization (no visualization vs. 2 × 2 table vs. tree diagram), (2) information format (probabilities vs. frequencies), and (3) problem context (mammography vs. economics problem). All versions begin with a cover story (see also **Table 2**); after that, one of three different kinds of visualization (including no visualization) was given (**Figure 2**). Finally, the question was provided in the same format as the information in the text.

The correct solution for the mammography problem is 80 out of 1,030 (about 7.8%), and for the economics problem 208 out of 613 (33.9%). Note that the corresponding algorithm to calculate the Bayesian posterior probability is identical for 2 × 2 tables concerning both information formats. However, the algorithm for computing P(B|M+) based on a tree diagram differs substantially with respect to both information formats.

A response has been classified as a correct "Bayesian answer" if the exact probability or frequency solution was provided, or the probability solution was rounded up or down to the next full percentage point (e.g., in the mammography problem the correct solution is 7.8%, therefore answers between 7 and 8% were classified as a correct solution; see also Gigerenzer and Hoffrage, 1995).

Participants

The participants were $N = 259$ German secondary school students age 16–18. Students were recruited from 12 different classes (grade 11) at two Bavarian Gymnasiums. Note that in Germany there are different kinds of secondary school tracks. In order to study at a university, the Gymnasium (academic track) must be pursued. All students were familiar with 2 × 2 tables

and tree diagrams containing probabilities and with 2 × 2 tables containing frequencies but not with natural frequency trees.

The study was carried out in accordance with the University Research Ethics Standards. The principals of both schools approved conduction of the study (this is mandatory in Germany when testing school students). When conducting the study we did not collect personal data (our questionare did not include questions with regard to age, gender etc.). Students were informed that their participation was voluntary (two students refrained from participating) and anonymity was guaranteed. After the study participants were debriefed.

Results

Our study showed three important findings (**Figure 3**). First, students' performance was higher when information in the problems was presented in natural frequencies (42% correct inferences, averaged across context and visualization) instead of probabilities (5%), which supports our hypothesis 1. This finding holds when only problems *without* visualizations are compared (26% correct inferences in natural frequency versions vs. 2% correct inferences in probability versions, averaged across both contexts, which replicates previous findings, e.g., Gigerenzer and Hoffrage, 1995; Siegrist and Keller, 2011) and when problems *with* visualizations are compared (51% correct inferences in natural frequency versions vs. 6% correct inferences in probability versions, averaged across both contexts).

Second, the additional presentation of visualizations helps understanding (hypothesis 2): Averaging across all versions *with* visualization yields higher performance (28%) than averaging across all versions *without* visualizations (14%). Note that this difference is much stronger in the natural frequency versions (51% vs. 26%, averaged across both contexts) than in the probability versions (6% vs. 2%, see **Figure 2**). The fact that probability visualizations only have very limited effect is irritating since these visual aids are frequently applied in statistical text books (see Discussion).

Furthermore, participants showed better performance in almost every version of the economics problem (30% correct inferences, averaged across format of information and visualization) compared to the respective versions of the mammography problem (16%). Possible reasons will be debated in Section "Discussion."

In order to analyze the impact of information format and visualization simultaneously we ran binary logistic regressions. Since we had no hypothesis on possible effects of problem context we performed two logistic regressions for the mammography problem and for the economics problem separately. The independent variables were visualization (only distinguishing between *no visualization* vs. *visualization*) and information format, respectively. The dependent variable was the correctness of the solution (1 – correct solution, 0 – incorrect solution). The results of the statistical analyses are illustrated in **Table 3**. For both contexts model 1 shows the impact of information format, whereas model 2 shows the impact of information format and visualization simultaneously.

In both problem contexts we found significant coefficients regarding information format (hypothesis 1) and visualization (vs. no visualization; hypothesis 2). Additional analyses revealed no statistical differences between 2 × 2 table and tree diagram in each information format. Although **Figure 3** suggests a possible interaction of format and visualization the regression does not yield a respective significant coefficient. Note that the seeming interaction between format and visualization may be due to the floor effect with respect to the probability versions. However, considering **Figure 2** it becomes clear that visualizations of the numerical values in probability versions do not help substantially.

Discussion

According to general theories of information encoding and processing (e.g., Cognitive Load Theory, Sweller, 2003; Cognitive Theory of Multimedia Learning, Mayer, 2005), understanding of statistical information could be supported by presenting additional visual aids. In our study, participants' performance in two Bayesian reasoning tasks was higher when additionally 2 × 2 tables and tree diagrams containing natural frequencies were presented. However, when applying these visual aids for Bayesian inferences, the information format should be taken into account: both tools have only very limited effects when probabilities are included. Since in statistics text books and school curricula both probability visualizations – but not frequency trees – commonly are applied in order to foster insight, this finding is quite remarkable.

In general, our results are in line with the "frequentist hypothesis" (Gigerenzer and Hoffrage, 1995; Cosmides and Tooby, 1996) as well as the "nested sets hypothesis" (Barbey and Sloman, 2007). Regarding all problem versions, natural frequency versions resulted in higher performance levels compared to the respective probability versions. The low performance, however, in the natural frequency version of the mammography problem without visualization indicates only moderate statistical literacy in the participants of our study. Interestingly, the performance in the economics problem was much better than in the mammography problem under almost every condition. A possible reason might be the extreme base rate (1%) in the mammography problem which basically constitutes the cognitive illusion (in contrast, the result of the economics problem is no longer counterintuitive). Another reason might be that the context of the economics problem is more adapted to the living environment of young people (a strong dependency from the problem context was also found by Siegrist and Keller, 2011). The more complicated terminology or taxing cognitive capacity in the mammography problem could also account for the deviant effects in the different contexts (e.g., Lesage et al., 2013; Sirota et al., 2014a).

The need for tools for teaching statistics is repeatedly stressed (Gigerenzer, 2013, 2014; Navarrete et al., 2014). There are several teaching studies (Sedlmeier and Gigerenzer, 2001; Wassner, 2004; Mandel, 2015; Sirota et al., 2015b) where the solution process of a Bayesian reasoning problem is explained explicitly, e.g.,

with the help of visualizations, and the effect of teaching is investigated. For instance, it is even possible to advise students to imagine an arbitrary sample when given a probability version and then to construct a frequency table or tree diagram accordingly (by increasing the size of the arbitrary sample whole numbers always can be reached for each respective subset). Furthermore Hoffrage et al. (submitted, same issue) instructed participants to solve complex Bayesian reasoning problems (e.g., with more than one cue) by translating the given information in terms of probabilities into natural frequencies and to construct a corresponding tree diagram accordingly. Note again, that our study is not an explicit teaching study; nevertheless our findings have pragmatic implications for teaching Bayesian reasoning. Our visualizations have the advantage that they can be constructed easily by teachers or students. In contrast, the diagrams in **Figure 1** are complicated to produce, which is especially problematic when base rates are extreme. In the unit square, for instance, areas can become very small (in **Figure 1** therefore a higher base rate of the disease was chosen). Similarly, concerning the icon array, more symbols would be required in the case of small or unmanageable proportions (such as 1.25 or 9.6%) thus entailing an enormous effort. Our frequency visualizations, which of course can be combined with other visualizations (for an integration of a natural frequency tree and an icon array see, e.g., Mossburger, unpublished manuscript), thus may be a helpful aid for fostering statistical understanding and for teaching statistics in schools.

Note that 2 × 2 tables and tree diagrams containing natural frequencies can not only aid in Bayesian reasoning problems, but can also illustrate situations with two dichotomous features in general. For instance, it is possible to justify and explain the rules for multiplication and addition of conditional probabilities with natural frequency trees very easily (Mossburger, unpublished manuscript). Since 2 × 2 tables and tree diagrams containing natural frequencies can be provided long before students have to solve Bayesian reasoning problems, these visual aids offer the opportunity to consider various types of problems over a long period of a school or university curriculum.

Acknowledgments

We thank both reviewers for helpful comments and Robert DeHaney for editing of an earlier version of the manuscript. This work was supported by the German Research Foundation (DFG) within the funding program Open Access Publishing.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *J. Pers. Soc. Psychol* 35, 303–314. doi: 10.1037/00223514.35.5.303

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–297. doi: 10.1017/S0140525X07001653

Bea, W. (1995). *Stochastisches Denken [Stochastical Reasoning]*. Frankfurt am Main: Peter Lang.

Brase, G. L. (2008). Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol.* 23, 369–381. doi: 10.1002/acp.1460

Brase, G. L. (2014). The power of representation and interpretation: doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *J. Cogn. Psychol.* 26, 81–97. doi: 10.1080/20445911.2013.861840

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

Dougherty, M. R., Gettys, C. F., and Ogden, E. E. (1999). MINERVA-DM: a memory processes model for judgments of likelihood. *Psychol. Rev.* 106, 180–209. doi: 10.1037/0033-295X.106.1.180

Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine: problems and opportunities," in *Judgment under Uncertainty: Heuristics and Biases*, eds D. Kahneman, P. Slovic, and A. Tversky (New York: Cambridge University Press), 249–267. doi: 10.1017/CBO9780511 809477.019

Ellis, K. M., Cokely, E. T., Ghazal, S., and Garcia-Retamero, R. (2014). Do people understand their home HIV test results? Risk literacy and information search. *Proc. Hum. Fact. Ergon. Soc. Annu. Meet.* 58, 1323–1327. doi: 10.1177/1541931214581276

Fiedler, K., Brinkmann, B., Betsch, T., and Wild, B. (2000). A sampling approach to biases in conditional probability judgments: beyond base rate neglect and statistical format. *J. Exp. Psychol.* 129, 399–418. doi: 10.1037/0096-3445.129.3.399

Friederichs, H., Ligges, S., and Weissenstein, A. (2014). Using tree diagrams without numerical values in addition to relative numbers improves students' numeracy skills: a randomized study in medical education. *Med. Decis. Making* 34, 253–257. doi: 10.1177/0272989X13504499

Garcia-Retamero, R., Cokely, E. T., and Hoffrage, U. (2015). Visual aids improve diagnostic inferences and metacognitive judgment calibration. *Front. Psychol.* 6:932. doi: 10.3389/fpsyg.2015.00932

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

Gigerenzer, G. (2013). HIV screening: helping clinicians make sense of test results to patients. *BMJ* 347, f5151. doi: 10.1136/bmj.f5151

Gigerenzer, G. (2014). How I got started: teaching physicians and judges risk literacy. *Appl. Cogn. Psychol.* 28, 612–614. doi: 10.1002/acp.2980

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033295X.102.4.684

Gigerenzer, G., and Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev.* 106, 425–430. doi: 10.1037/0033-295X.106.2.425

Girotto, V., and Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition* 78, 247–276. doi: 10.1016/S0010-0277(00)00133-5

Girotto, V., and Gonzalez, M. (2002). Chances and frequencies in probabilistic reasoning: rejoinder to Hoffrage, Gigerenzer, Krauss, and Martignon. *Cognition* 84, 353–359. doi: 10.1016/S0010-0277(02)00051-3

Goodie, A. S., and Fantino, E. (1996). Learning to commit or avoid the base-rate error. *Nature* 380, 247–249. doi: 10.1038/380247a0

Hoffrage, U., and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Acad. Med.* 73, 538–540. doi: 10.1097/00001888-199805000-00024

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition* 84, 343–352. doi: 10.1016/S0010-0277(02)00050-1

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science* 290, 2261–2262. doi: 10.1126/science.290.5500.2261

Kahneman, D., and Frederick, S. (2005). "A model of heuristic judgment," in *The Cambrigde Handbook of Thinking and Reasoning*, eds K. J. Holyoak and R. G. Morris (Cambridge: Cambridge University Press), 267–293.

Kleiter, G. D. (1994). "Natural sampling: rationality without base rates," in *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, eds

G. H. Fischer and D. Laming (New York: Springer), 375–388. doi: 10.1007/978-1-4612-4308-3_27

Krauss, S., and Bruckmaier, G. (2014). "Eignet sich die Formel von Bayes für Gerichtsverfahren? [Is formula of Bayes appropriate for legal trials?]," in *Daten, Zufall und der Rest der Welt*, eds U. Sproesser, S. Wessolowski, and C. Wörn (Wiesbaden: Springer), 123–132. doi: 10. 1007/978-3-658-04669-9_10

Lesage, E., Navarrete, G., and De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: the role of general cognitive resources. *Think. Reason.* 19, 27–53. doi: 10.1080/13546783.2012.713177

Lewis, C., and Keren, G. (1999). On the difficulties underlying Bayesian reasoning: a comment on Gigerenzer and Hoffrage. *Psychol. Rev.* 106, 411–416. doi: 10.1037/0033-295X.106.2.411

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Front. Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Front. Psychol.* 6:387. doi: 10.3389/fpsyg.2015.00387

Mayer, R. E. (2005). "Cognitive theory of multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, ed. R. E. Mayer (New York: Cambridge University Press), 31–48. doi: 10.1017/CBO9780511816819.004

Mellers, B. A., and McGraw, A. P. (1999). How to improve Bayesian reasoning: comment on Gigerenzer and Hoffrage (1995). *Psychol. Rev.* 106, 417–424. doi: 10.1037/0033-295X.106.2.417

Micallef, L., Dragicevic, P., and Fekete, J. (2012). Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. *IEEE Trans. Vis. Comput. Graph.* 18, 2536–2545. doi: 10.1109/TVCG.2012.199

Navarrete, G., Correia, R., and Froimovitch, D. (2014). Communicating risk in prenatal screening: the consequences of Bayesian misapprehension. *Front. Psychol.* 5:1272. doi: 10.3389/fpsyg.2014.01272

Paling, J. (2003). Strategies to help patients understand risks. *BMJ* 327, 745–748. doi: 10.1136/bmj.327.7417.745

Ruscio, J. (2003). Comparing Bayes's theorem to frequency-based approaches to teaching Bayesian reasoning. *Teach. Psychol.* 30, 325–328.

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Siegrist, M., and Keller, C. (2011). Natural frequencies and Bayesian reasoning: the impact of formal education and problem context. *J. Risk Res.* 14, 1039–1055. doi: 10.1080/13669877.2011.571786

Sirota, M., Juanchich, M., and Hagmayer, Y. (2014a). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychonom. Bull. Rev.* 21, 198–204. doi: 10.3758/s13423-013-0464-6

Sirota, M., Kostovičová, L., and Juanchich, M. (2014b). The effect of iconicity of visual displays on statistical reasoning: evidence in favor of the null hypothesis. *Psychon. Bull. Rev.* 21, 961–968. doi: 10.3758/s13423-013-0555-4

Sirota, M., Kostovičová, L., and Vallée-Tourangeau, F. (2015a). Now you Bayes, now you don't: effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychon. Bull. Rev.* doi: 10.3758/s13423-015-0810-y [Epub ahead of print].

Sirota, M., Kostovičová, L., and Vallée-Tourangeau, F. (2015b). How to train your Bayesian: a problem-representation transfer rather than a format-representation shift explains training effects. *Q. J. Exp. Psychol.* 68, 1–9. doi: 10.1080/17470218.2014.972420

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychol. Bull.* 119, 3–22. doi: 10.1037/0033-2909.119.1.3

Sloman, S. A., Over, D., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Process* 91, 296–309. doi: 10.1016/S0749-5978(03)00021-9

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science* 333, 1393–1400. doi: 10.1126/science.1191181

Sturm, A., and Eichler, A. (2014). "Students' beliefs about the benefit of statistical knowledge when perceiving information through daily media," in *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9), Flagstaff, AZ: Sustainability in Statistics Education*, eds K. Makar, B. de Sousa, and R. Gould (Voorburg: International Statistical Institute.).

Sweller, J. (2003). Evolution of human cognitive architecture. *Psychol. Learn. Motiv.* 43, 215–266. doi: 10.1145/1404520.1404521

Wassner, C. (2004). *Förderung Bayesianischen Denkens – Kognitionspsychologische Grundlagen und didaktische Analysen [Promoting Bayesian Reasoning – Principles of Cognitive Psychology, and Didactical Analyses]*. Hildesheim: Franzbecker.

Zikmund-Fisher, B. J., Witteman, H. O., Dickson, M., Fuhrel-Forbis, A., Kahn, V. C., Exe, N. L., et al. (2014). Blocks, ovals, or people? Icon type affects risk perceptions and recall of pictographs. *Med. Decis. Making* 34, 443–453. doi: 10.1177/0272989X13511706

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 22 June 2015
doi: 10.3389/fpsyg.2015.00642

Natural frequencies facilitate diagnostic inferences of managers

Ulrich Hoffrage[1]*, *Sebastian Hafenbrädl*[1] and *Cyril Bouquet*[2]

[1] Department of Organizational Behavior, Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland,
[2] International Institute for Management Development, Lausanne, Switzerland

In Bayesian inference tasks, information about base rates as well as hit rate and false-alarm rate needs to be integrated according to Bayes' rule after the result of a diagnostic test became known. Numerous studies have found that presenting information in a Bayesian inference task in terms of natural frequencies leads to better performance compared to variants with information presented in terms of probabilities or percentages. Natural frequencies are the tallies in a natural sample in which hit rate and false-alarm rate are not normalized with respect to base rates. The present research replicates the beneficial effect of natural frequencies with four tasks from the domain of management, and with management students as well as experienced executives as participants. The percentage of Bayesian responses was almost twice as high when information was presented in natural frequencies compared to a presentation in terms of percentages. In contrast to most tasks previously studied, the majority of numerical responses were lower than the Bayesian solutions. Having heard of Bayes' rule prior to the study did not affect Bayesian performance. An implication of our work is that textbooks explaining Bayes' rule should teach how to represent information in terms of natural frequencies instead of how to plug probabilities or percentages into a formula.

Keywords: bayesian inference, updating beliefs, natural frequency, representation format, management, executives, applied business statistics

OPEN ACCESS

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
Gary L. Brase, Kansas State University, USA
Elisabet Tubau, Universitat de Barcelona, Spain

***Correspondence:**
Ulrich Hoffrage, Department of Organizational Behavior, Faculty of Business and Economics, University of Lausanne, Batiment Internef, CH-1015 Lausanne-Dorigny, Switzerland
ulrich.hoffrage@unil.ch

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *17 February 2015*
Accepted: *01 May 2015*
Published: *22 June 2015*

Citation:
Hoffrage U, Hafenbrädl S and Bouquet C (2015) Natural frequencies facilitate diagnostic inferences of managers. Front. Psychol. 6:642. doi: 10.3389/fpsyg.2015.00642

Introduction

Twenty years ago, Gigerenzer and Hoffrage (1995) demonstrated that Bayesian inferences can be improved without instructing participants how to solve such Bayesian tasks. By providing the relevant information not in terms of probabilities, percentages, or relative frequencies, as it is usually done, but in terms of natural frequencies, the percentage of correct (i.e., Bayesian) inferences tripled, specifically, from 16 to 46%. What is a Bayesian inference task and what are natural frequencies? Consider the following example:

> *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 material 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?*

If the question was "*What is the probability that material, randomly drawn from the inventory, comes from supplier A*," then the answer would be easy: 30%. Since 30 and 70% are the base rates for the

two suppliers, A and B, respectively, one could simply use these base rates when asked about the prior probability that material comes from suppliers A or B. Taking supplier A as a reference, these two probabilities will henceforth be referred to as *p(H)* and *p(−H)*, which is the standard notation for the probability that a hypothesized event will occur (or not), or whether a hypothesis is true (or not).

The term "prior" refers to the point in time before diagnostic information has been given. In the example above, such data (D) has indeed been observed—specifically, the material has been identified as defective. This information should be used to update the prior probability. Following this update, the best estimate that the material comes from supplier A is the posterior probability, *p(H|D)*. It can be calculated using Bayes' rule:

$$p(H|D) = \frac{p(H)p(D|H)}{p(H)p(D|H) + p(-H)p(D| - H)} \quad (1)$$

where *p(D|H)* stands for the probability that material is defective if it comes from supplier A (in the example above, this probability is given by the relative frequency of 15%), and where *p(D|−H)* stands for the probability that material is defective if it comes from supplier B (in the example above, 10%).

Previous research has shown that people have difficulties to infer the posterior probability from the prior probability and the two likelihoods, *p(D|H)* and *p(D|−H)* (in terms of signal-detection theory, these two likelihoods are referred to as hit rate and false-alarm rate, respectively; in medical terms, the hit rate is called sensitivity and the false-alarm rate is the complement of the specificity). In order to give the reader a better chance to experience some empathy with participants, we do not reveal the Bayesian solution to the Skiwell Manufacturing Company task at this point—but note that the task was even harder for the participants because they, unlike the reader, did not have Equation 1 at their disposal. Kahneman and Tversky (1972) concluded from their research that participants do not integrate the three pieces of information; they rather confuse the posterior probability, *p(H|D)*, with the likelihood of the observed data if the prior hypothesis were true, *p(D|H)*, and provide the latter as an answer when asked for the former. Kahneman and Tversky consider this confusion as an application of the representativeness heuristic—which Gigerenzer (1996), in turn, considers to be a re-description or a "one-word explanation" (p. 594; see also Gigerenzer and Murray, 1987). Using the representativeness heuristic amounts to ignoring the base rates, which Kahneman and Tversky (1972) demonstrated with a between-subjects design: The posterior beliefs of two groups of participants were indistinguishable even though these two groups received different base rates and should hence have different prior probabilities. The authors concluded that "In his evaluation of evidence man is apparently not a conservative Bayesian: he is not Bayesian at all" (p. 450). This "base-rate neglect" is one of the prime examples for a cognitive fallacy investigated in the "heuristics and biases" program (Kahneman et al., 1982), and Bar-Hillel (1980) stated that "the genuineness, the robustness, and the generality of the base-rate fallacy are matters of established fact" (p. 215).

This conclusion has been challenged by Gigerenzer and Hoffrage (1995) with a study in which they represented the information about base rate and the two likelihoods in terms of natural frequencies. Using this representation format, our task reads as follows:

> *The Skiwell Manufacturing Company gets material from two suppliers. Out of 1,000 items, supplier A delivers 300 and supplier B delivers the remaining ones. Past records indicate that 45 of the 300 items delivered by supplier A are defective and that 70 out of the 700 items delivered by B 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: How many of the items that have been identified as defective come from supplier A?*

Natural frequencies are the frequencies that naturally result if a sample is taken from a population (or if the entire population is considered). In case of one hypothesis (H, with its complement −H) and one dichotomous, diagnostic variable that represents the data (D), natural frequencies are the four entries in the bivariate 2×2 table. The frequencies of the four conjunctive events can be displayed in two trees, in each of which the total sample size (or population) is the top node, and the four possible combinations are on the lowest level. One of these two possible trees displays the row margins at the intermediate level, and the other one the column margins. For instance, in **Figure 1**, Panel B, the two natural frequencies for a sample of 1,000 items are displayed at the intermediate level: 300 come from supplier A and 700 from supplier B, corresponding to the two base rates of 30 and 70%. From this tree in Panel B, it is relatively easy to determine the total number of defective items ($45 + 70 = 115$), and the total number of intact items ($255 + 630 = 885$). These two numbers are basically the margins of the diagnostic variable, and the first is included in the Bayesian solution to our task: Of the 115 defective items, 45 were delivered by supplier A. This is also the Bayesian response that we withheld above when we presented the problem in terms of percentages: $p(H|D) = 0.39$ (or, as a ratio, 45/115). From **Figure 1**, Panel B, it is also easy to construct the tree displayed in Panel C, which would also allow one to answer to other questions, for instance, how many of the intact items were delivered by supplier B.

The tree in Panel A displays the information as it has been represented in the initial version of the Skiwell Manufacturing Company task. What made it hard to derive the solution from this representation, compared to a natural frequency representation, was the fact that the two likelihoods have been normalized with respect to the base rates, and for exactly this reason, Gigerenzer and Hoffrage (1995) predicted that representations in terms of probabilities, percentages, and relative frequencies will not differ with respect to Bayesian performance (Prediction 4, p. 692). For a more detailed discussion of the notion of natural frequencies and its relationship to other representation formats, see Hoffrage et al. (2002), Gigerenzer and Hoffrage (2007), and Johnson and Tubau (in review).

Natural Frequencies have proven to facilitate diagnostic inferences in laypeople (Gigerenzer and Hoffrage, 1995), advanced medical students and advanced law students (Hoffrage

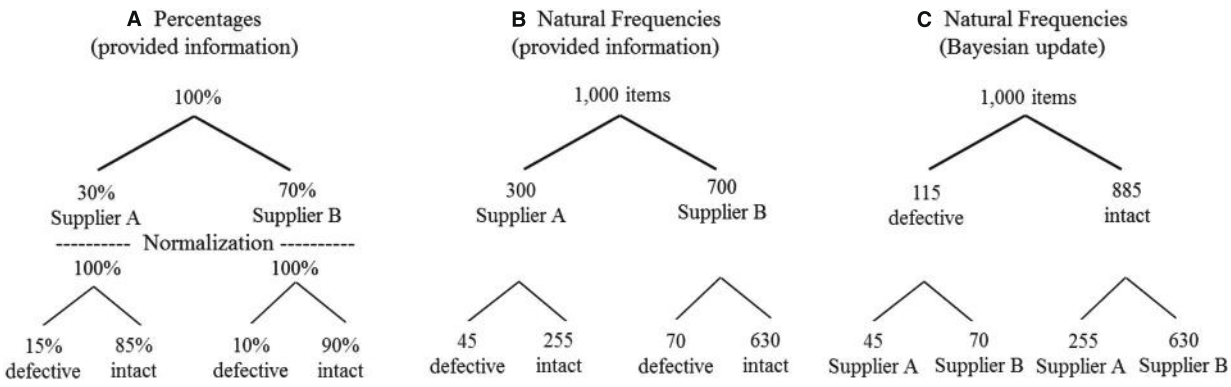

FIGURE 1 | Numerical information of the Skiwell Manufacturing Company task. (A) Information provided in percentages. Hit rate and false-alarm rate have been normalized with respect to the base rates of the two suppliers. **(B)** Natural frequencies with suppliers at the intermediate level. The frequencies of the four conjunctive events implicitly contain the base rate information about the suppliers. **(C)** Natural frequencies with diagnostic information at the intermediate level. From the perspective of **(B)**, the tree in **(C)** represents a Bayesian update, in which the four distinct events are now conditioned on diagnostic information.

et al., 2000), patients (Garcia-Retamero and Hoffrage, 2013), and physicians (Hoffrage and Gigerenzer, 1998). This result is well established (Mandel, 2015), it has been replicated by many others (e.g., Akl et al., 2011; Woloshin and Schwartz, 2011), and this work has received wide attention in the medical field and beyond (Gigerenzer, 2002, 2014; Gigerenzer et al., 2007; Gigerenzer and Gray, 2011). For a discussion about when and why natural frequencies are effective, see Brase (2008), Brase and Hill (2015), Gigerenzer and Hoffrage (2007), Hill and Brase (2012), and Johnson and Tubau (in review).

Bayesian inference problems are also vital to management decisions. For instance, a sales manager may be interested in whether a customer places more weight on quality than price if her yearly income is above average, a bank may be interested in whether it will see the annuity for a mortgage if the customer will lose his job, a project manager may be interested in whether the group will be able to complete the project in time if one of the key engineers will get sick unexpectedly, and so on. The fact that the task of updating beliefs is ubiquitous and also relevant in the world of business makes it even more surprising that, to the best of our knowledge, there is no research investigating whether natural frequencies are also beneficial for managers and management problems. This is exactly the aim of the present paper. The participants in the studies reported below were executives and business students who had to work on four different tasks with business-related content.

Materials and Methods

Participants

Participants were undergraduates at a business faculty ($n = 259$) and executives ($n = 181$; for a total n of 440). The undergraduates were either students enrolled in their third year of the Bachelor of Science in Management program of a public Swiss university who took the lecture "Judgment and Decision Making" of the first author, or students enrolled in their first year of the Master of Science in Management program who took the seminar "Analytic and Intuitive Judgment" of the second author. Over 3 years, three cohorts of bachelor students and in the fourth year, one cohort of bachelor students and one cohort of master students were tested, with 74, 45, 49, 62, and 29 students responding to the questionnaire. Demographic information was only collected for the last two cohorts: The bachelor students were on average 21.4 years old (SD = 1.2) and 51% were female, and the master students were on average 24 years old (SD = 1.9) and 30% were female. The entire population of the three earlier cohorts of bachelor students was demographically similar to the bachelor students of the last cohort.

The executives were also tested in a classroom setting, namely in their role as students in an executive MBA program. In fact, they enrolled in either of two different programs. One was the Executive MBA program offered by a public Swiss university in which they took a course "Managerial Decision Making and Negotiation" of the first author. Four different cohorts from four different years have been tested, with 27, 28, 22, and 43 respondents ($n = 120$). Average age was 38, 38, 37, and 38 years, and 87, 83, 72, and 76% were male. These participants are henceforth referred to as junior managers. The other program (Program for Executive Development, in fact a very prestigious and competitive program) was offered by a private Swiss business school. The executives took a course taught by the third author who had invited the first author as a guest lecturer. Two cohorts of the same module have been tested, with 27 and 34 participants each ($n = 61$). With an average age of 42 years, these executives were older than the ones from the public university, and so they are henceforth referred to as senior managers. In fact, many of them were directors or vice-presidents in their companies.

Materials, Design, and Procedure

Four tasks have been used, all adapted from Groebner et al. (2007). The Skiwell Manufacturing task introduced above was one of them, the three others involved error/fraud detection (IRS Audit), success in the context of an auction (Techtronics Equipment), and quality control (Varden Soap). For each task,

TABLE 1 | The four tasks used in the present study with the information provided and the Bayesian solution.

Task	Condition	Base rate	Hit rate	False-alarm rate	Bayesian solution
Skiwell manufacturing	Percentages	30	15	10	39.13
	Natural frequencies	300 of 1000	45 of 300	70 of 700	45 of 115
IRS Audit	Percentages	20	30	10	42.86
	Natural frequencies	200 of 1000	60 of 200	80 of 800	60 of 140
Techtronics equipment	Percentages	60	70	50	67.74
	Natural frequencies	60 of 100	42 of 60	20 of 40	42 of 62
Varden soap	Percentages	60	5	10	42.86
	Natural frequencies	600 of 1000	30 of 600	40 of 400	30 of 70

two versions were constructed, one in which the information was presented in percentages and one in which natural frequencies were used (see the Appendix for the exact formulations of these other three tasks, and **Table 1** for the numbers involved in all four tasks).

Each questionnaire consisted of two different tasks, either two percentage versions, or two natural frequency versions. Which task was paired with which other task and their order within the same questionnaires was counterbalanced, so that the number of respondents per task and per version was, ideally, equally distributed (minor deviations from an equal distribution were due to the fact that the number of students in a classroom was rarely divisible by the minimal number of questionnaires that would allow for an equal distribution, resulting in 111, 110, 111, 110 for the percentage versions, and 110, 108, 111, 109 for the frequency versions of Skiwell, IRS, Techtronics, and Varden Soap, respectively).

Students were given 7 min to work on the tasks. They were allowed to take notes. Some students had a pocket calculator with them (or a smartphone with this function), and very few asked whether they could use them. The answer was positive, but with respect to those who did not have one at their disposal, it was added that writing down a mathematical operation, for instance, a ratio, would be sufficient. In other words, we made it clear that we were not interested in whether they could enter numbers into a pocket calculator, but whether they were able to figure out *which* numbers to enter, and that writing down the correct operation would be treated as a correct response even if they would not convert it into an exact decimal. After 7 min the questionnaires were collected, but it could not have been prevented that some students continued writing during the collection procedure.

This procedure was slightly altered for the 62 bachelor students and the 29 master students who were tested during the last year of data collection. After 7 min, they were prompted to turn the questionnaire to a new page that was not included in the questionnaire of the 181 executives and the other 168 undergraduates. On this page, they entered their demographics and responded to several questions concerning their prior knowledge about Bayes' rule. To prevent participants from being exposed to the term "Bayes' rule" within the questionnaire before finishing the inference problems, these questions were only displayed via a projector once all students had finished working on the inference problems.

After having turned in their questionnaires, students received a lecture about Bayesian inferences and representation formats—sometimes right after the questionnaires, sometimes in another lecture. When those participants whose booklet did *not* include the page with the questions regarding Bayes' rule were asked, during this debrief, whether they were familiar with this kind of task and whether they had received some instructions or training beforehand, for instance, in a lecture on statistics, very few (about 5% of the executives and about 10% of the management undergraduates) raised their hand.

Analysis

The analysis was mainly based on outcomes, that is, on participants' numerical responses. Following Gigerenzer and Hoffrage (1995), a response has been classified as Bayesian if the absolute difference between this response and the Bayesian solution was lower than one percentage point. This criterion was lenient enough to also include rounding up or down to the next whole number. In fact, many participants in the percentage condition were able to derive the Bayesian solution, wrote down the formula, used the pocket calculator of their smartphone to compute the exact value, but then wrote, for the Skiwell Manufacturing task, 39 or 40% (instead of 39.1304%).

Traces of cognitive processes, that is, notes and remarks that revealed how participants arrived at their answers, were also considered. If a participant provided a numerical response that we would have classified as Bayesian, but if the notes made it clear that this match was only coincidental and resulted from a non-Bayesian rationale, then we did not classify the response as Bayesian. Conversely, if a participant in the percentage condition wrote down a ratio that corresponded to the Bayesian solution, but did not compute the exact number (by hand or with a pocket calculator), we nevertheless classified it as a Bayesian answer—and as we already mentioned above, participants were informed about this.

Results

Do Natural Frequencies Facilitate Bayesian Inferences in Our Four Tasks?

Yes. **Figure 2** displays the percentages of Bayesian responses, both for the four tasks separately and across all tasks. In the percentage condition, 87 of 442 responses (19.7%) were Bayesian, and in the natural frequency condition, these were 170 of 438

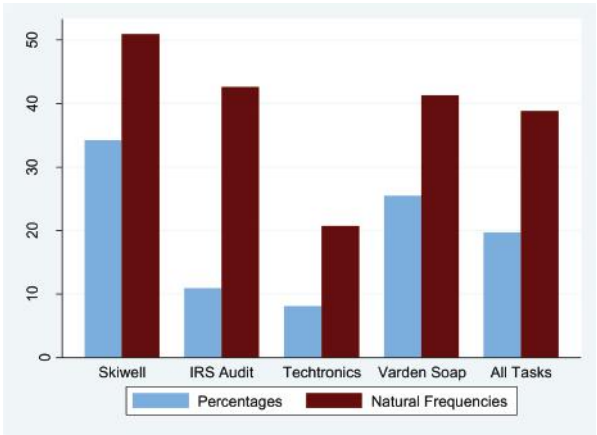

FIGURE 2 | Percentages of Bayesian responses, depending on whether the numerical information has been communicated in terms of percentages or natural frequencies.

(38.8%). A logistic regression confirmed that the format in which information was presented had a significant effect ($B = 1.04$, SE $= 0.20$, $z = 5.24$, $p < 0.001$; after controlling for task, order, and participant sample, and with standard errors clustered for each participant).

Does Representation Format also Affect the Non-Bayesian Inferences?

While the previous analysis focused on the percentages of Bayesian responses, it is also useful to take a look at the full distribution of numerical estimates, independent of whether participants succeeded in deriving the Bayesian solution or not. **Figure 3** provides such a more fine-grained picture of the distribution of numerical estimates for the four tasks. Two estimates in the natural frequency condition for which the numerator was larger than the denominator, and one of 125% in the percentage condition were classified as non-Bayesian in **Figure 2**, but were not graphically displayed in **Figure 3**.

Overall, most estimates were too low: Across all tasks in the percentage condition, 58.4% of the responses were lower than the Bayesian solution, 19.7% were classified as Bayesian, 21.7% were higher than the Bayesian solution and 0.2% were above 100%. For the natural frequency condition, these numbers were 43.4, 38.8, 17.4, and 0.4% respectively. When information has been presented in terms of natural frequencies, the responses were not only more often correct (see **Figure 2**), but also closer to the Bayesian solution: The average absolute difference between responses and Bayesian solution was 19.2 in the percentage condition, and 15.1 for natural frequencies (excluding responses above 100%). Regression analysis revealed that this difference was significant ($B = 3.84$, SE $= 1.35$, $t = 2.84$, $p = 0.005$; after controlling for task, order, and participant sample, and with standard errors clustered for each participant). Closer inspection, however, revealed that this difference was mainly due to the Bayesian responses. After these have been excluded, the picture even reversed: In the percentage condition, the average absolute difference of the remaining cases was 24.5, and in the natural frequency condition, it was 26.3 (but this effect was not significant: $B = 2.12$, SE $= 1.27$, $t = 1.68$, $p = 0.095$). In sum, in each of the two experimental conditions, most responses were too low. Participants' responses were closer to the Bayesian solution in the natural frequency condition, but this effect was mainly due to the fact that there were more Bayesian responses in the first place.

Figure 3 also shows that most numerical estimates were either identical to one of the pieces of information that has been given for a particular task, namely the base rate (Br), the hit rate (Hr), or the false-alarm rate (F), or that they matched the Bayesian response (Bay) or the probability that D and H occur together (joint occurrence, J, which is the product of hit rate times base rate of focal hypothesis). Results from a more detailed analysis of the most frequently used cognitive strategies—indicated by the lines in **Figure 3**—will be reported in the next section that focusses on the effects of participant sample (**Table 2**).

Who Performed Better: The Undergraduates or the Executives?

Figure 4 displays the percentages of Bayesian responses for the different types of participants. No clear picture emerged. While the undergraduates performed worse than the executives in the percentage condition (14.6, 28.6, 22.6, and 26.6%, for undergraduates, junior executives, senior executives, and executives combined, respectively), they outperformed the executives when the information was represented in natural frequencies (40.5, 39.5, 30.0, and 36.2%, respectively).

The main effect of participant sample was not significant ($B = 0.219$, SE $= 0.26$, $z = 0.86$, $p = 0.392$), but the interaction between participant sample and representation format was ($B = 0.99$, SE $= 0.40$, $z = 2.52$, $p = 0.012$; after controlling for representation format, task, and order, and with standard errors clustered for each participant). Analyzing the contrast between undergraduates and executives (junior and senior combined) separately, revealed a significant difference in the percentage condition (14.6 vs. 26.6%, $B = 0.796$, SE $= 0.310$, $z = 2.57$, $p = 0.010$), while the difference in the natural frequency condition was negligible and not significant (40.5 vs. 36.2%, $B = 0.213$, SE $= 0.252$, $z = 1.12$, $p = 0.398$). Finally, analyzing the contrast between the two representation formats revealed a significant difference between the percentage condition and the natural frequency condition within the undergraduate sample (14.6 vs. 40.5%, $B = 1.48$, SE $= 0.28$, $z = 5.32$, $p < 0.001$), while the superiority of natural frequencies did not reach statistical significance within the sample of executives (junior and senior combined; 26.6 vs. 36.2%, $B = 0.47$, SE $= 0.28$, $z = 1.67$, $p = 0.095$).

Table 2 completes the picture by also including the non-Bayesian strategies. While **Figure 3** splits the frequencies of the five different cognitive strategies according to representation format and task, **Table 2** splits them according to representation format and participant sample (again, undergraduates vs. executives; the latter with junior and senior combined). This table also displays, for each strategy separately, the coefficients (B) and the p-values of the five different logistic regressions, each with the

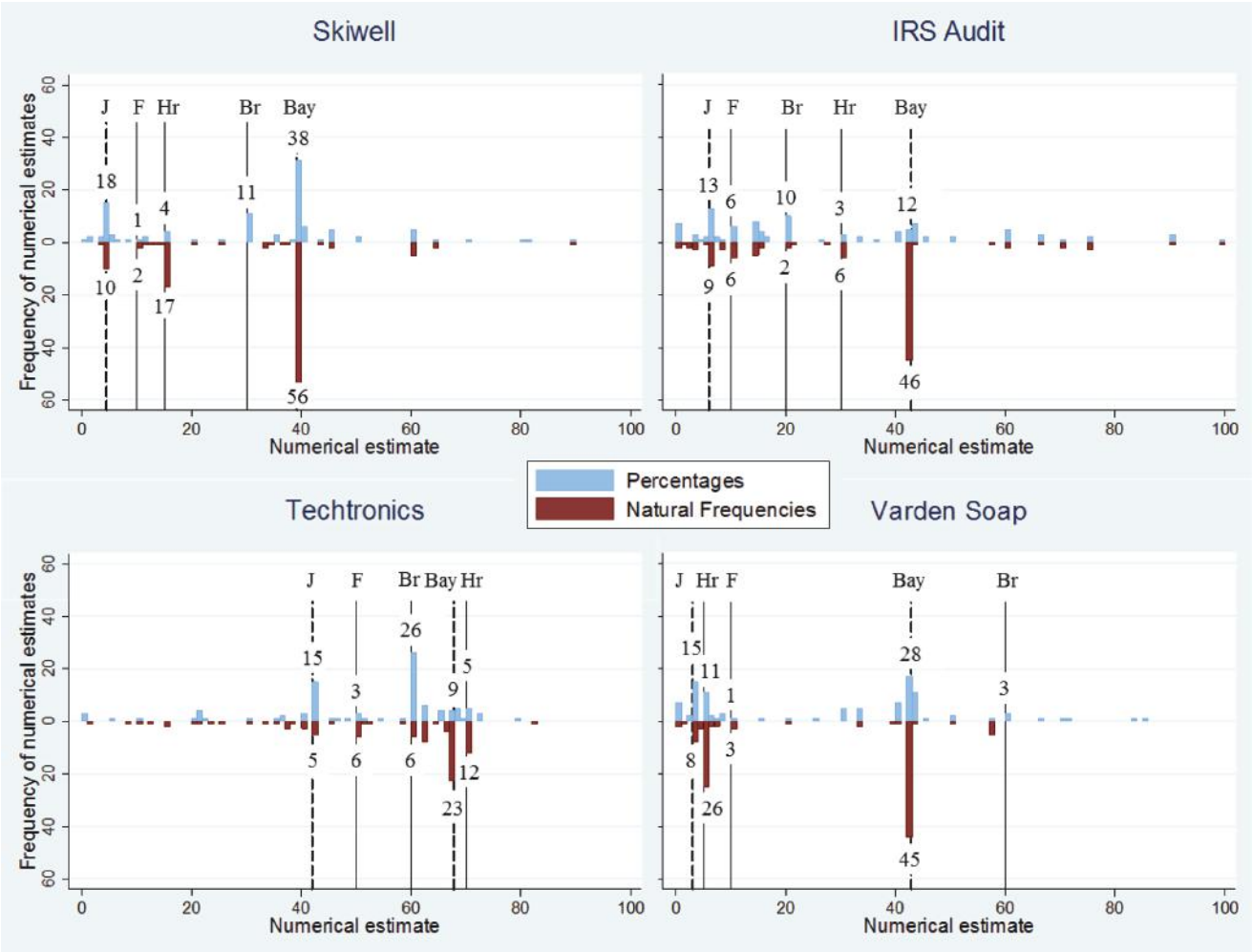

FIGURE 3 | Distribution of numerical estimates for the four tasks. The three straight lines indicate information that has been given in the task: Br indicates the base rate for the focal category, Hr indicates the hit rate (that is, diagnostic information conditioned on the focal category, $p(D|H)$), F indicates the false-alarm rate (that is, diagnostic information conditioned on the non-focal category, $p(D|-H)$). The two dotted lines indicate possible ways of combining this information: Bay indicates the Bayesian solution, p($H|D$), and J stands for Joint Occurrence of D and H, $p(D \text{ and } H)$. The numbers on these lines (and those on the y-axes) denote the frequencies of the corresponding numerical estimates. For instance, 10 participants in the percentage condition of the IRS task provided a numerical estimate between 20 and 20.99% (in fact, all these 10 participants wrote exactly 20%, which was identical to the base rate of that task), and 12 participants provided an estimate that has been coded as a Bayesian response (three gave the exact Bayesian response, either as the ratio 45/115, or they wrote down the exact number including the decimal, 42.86%, most likely with the help of a pocket calculator, one responded with 42.8%, and one with 42.9%. These five responses are displayed in the bracket ranging from 42.0–42.99%. The remaining seven participants responded with 43%. These seven estimates are displayed in the adjacent bracket, namely 43.0–43.99%, but they were nevertheless coded as Bayesian because our classification criterion allowed for rounding within one percentage point, see above). Note that something similar could be observed for each of the four tasks: the responses that have been classified as Bayesian are spread across two adjacent brackets, and hence the number of Bayesian responses is not visualized by one single bar, but rather consists of two lower numbers visualized by two bars.

main effect of representation format and participant sample, and the interaction between representation format and sample (after controlling for task and order, and with standard errors clustered for each participant). For each of the five cognitive strategies, except for providing the false-alarm rate as response, the number of participants who provided the corresponding numerical estimate significantly differed between the percentage condition and the natural frequency condition. In contrast, for none of the strategies, except for joint occurrence, we observed a significant effect of participant sample, and for none of the strategies, except for Bayesian, the interaction between representation format and participant sample reached significance.

Did the Order Matter?

No. Across all tasks, participants, and both representation formats, a response has been classified as Bayesian in 30.5% for tasks on the first page of the questionnaire and 28.0% for tasks on the second page (in a logistic regression, the

difference was not significant; $B = 0.131$, SE $= 0.115$, $z = 1.14$, $p = 0.254$; after controlling for representation format, participant sample and task, and with standard errors clustered for each participant).

Did Prior Knowledge of Bayes' Rule Make a Difference?

For none of our executive participants, but for 91 of our 259 undergraduate participants (62 bachelor and 29 master students) the booklet contained questions on demographics and on prior knowledge about Bayes' rule. A majority of 46 (74%) of the bachelor students and 17 (59%) of the master students responded that they have heard of Bayes' rule before this lecture (overall, 63 of 91 = 69%). With the exception of one bachelor student, all of those who had heard about Bayes' rule said it was taught to them in a course: 35 (56%) at school and 27 (44%) at the university (62/91 = 68%). A minority of 27 (44 %) of the bachelor students and 13 (45%) of the master students responded that they know when Bayes' rule is applicable, and 7 (11%) of the bachelor students and 11 (38%) of the master students (18/91 = 20%) were able to provide a short and correct explanation (we applied a very lenient criterion and coded answers such as "when conditional probabilities need to be computed" as correct). When asked whether they are able to formulate Bayes' rule, 21 (34%) of the bachelor students and 8 (28%) of the master students (29/91 = 32%) responded with yes, but only 10 (16%) of the bachelor students and 1 (3%) of the master students (11/91 = 12%) wrote down the correct formula. We should add that none of these 11 students reproduced our Equation 1 exactly, instead they all used an abridged version and wrote $p(H|D) = p(H)p(D|H)/p(D)$—which we coded as correct, despite of the fact that we cannot exclude the possibility that someone used a smartphone with internet access, and despite having doubts that someone who wrote down this abridged version was able to use it for our Bayesian tasks and to understand that the denominator, $p(D)$, amounts to $p(H)p(D|H) + p(-H)p(D|-H)$. These doubts lead straightforward to our next question: How have these differences between students been reflected in their ability to produce Bayesian responses?

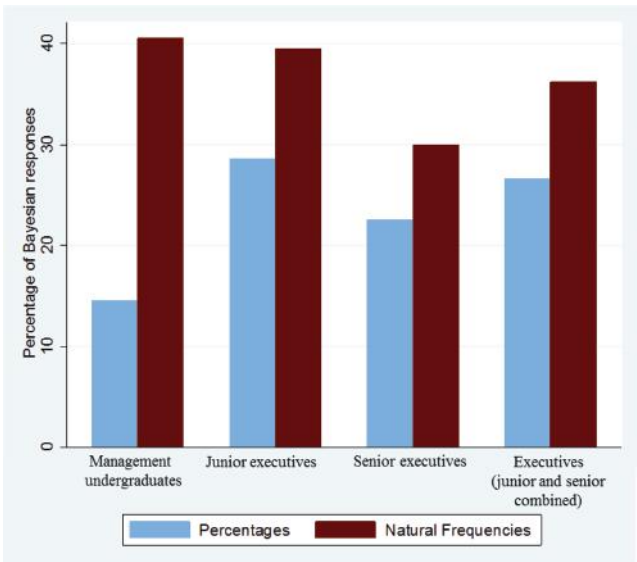

FIGURE 4 | Percentages of Bayesian responses, separately for the different groups of participants.

To the extent that teaching and instructions leave traces, one may expect that those who had heard of Bayes' rule performed better than those who had not. To test this hypothesis, we included the responses to each of the four questions about prior knowledge of Bayes' rule as a predictor of performance in a separate logistic regression, controlling for format, task, order, and participant sample, and with standard errors clustered for each participant. None of these regressions revealed a significant effect of prior knowledge on performance in our Bayesian tasks. To investigate whether prior knowledge affects the ability to produce Bayesian responses differentially, dependent on the format of the question, we reran the logistic regressions, this time with an additional interaction term between participants' responses and representation format. The result was the same: for none of the four questions concerning prior knowledge of Bayes' rule was there a significant main effect or a significant interaction effect on Bayesian performance.

TABLE 2 | Use of cognitive strategies, split by representation format and participant sample.

						Logistic regression results					
	Percentages		Natural frequencies			Format		Sample		Format × Sample	
Cognitive strategy	Undergraduates	Executives	Undergraduates	Executives	Total	*B*	*p*	*B*	*p*	*B*	*p*
Bayesian	14.57	26.6	40.53	36.21	29.2	1.47	<0.001	0.22	0.39	1.0	0.012
Base rate	11.42	11.17	1.52	2.3	6.6	−2.21	0.001	0.47	0.55	−0.47	0.587
Hit rate	4.72	5.85	13.26	14.94	9.5	1.15	0.001	0.12	0.69	0.05	0.886
False-alarm rate	1.97	3.19	4.17	3.45	3.2	0.78	0.174	0.18	0.76	0.7	0.413
Joint occurence	12.6	15.43	4.55	11.49	10.6	−1.11	0.009	1.0	0.02	−0.77	0.158
Total observations	254	188	264	174	880						

The coefficients (B) and p-values result from five different logistic regressions, one for each strategy, that were conducted to determine how representation format and participant sample affected strategy use (after controlling for task and order, and with standard errors clustered for each participant).
Total observations refer to the total number of responses on which the percentages reported in the cells are based, that is, the numbers in the cells denote column percentages.

General Discussion

To the best of our knowledge, the present study is the first to test whether natural frequencies facilitate Bayesian reasoning with management related tasks given to management undergraduates and executives. Even though the effect was not as strong as in previous studies, it is still larger than most effects observed in the social sciences: About twice as many participants came up with the Bayesian response when information was presented in terms of natural frequencies compared to percentages.

Distribution of Bayesian and Non-Bayesian Responses: Toward an Ecological Analysis of Bayesian Inference Tasks

A remarkable finding of our study is that most non-Bayesian estimates were lower than the Bayesian solution (**Figure 3**). This pattern is unusual, at least when compared to information in typical medical diagnostic tasks in which the Bayesian response is usually low and participants' responses are usually much higher (e.g., Eddy, 1982). How could one account for the different response patterns? One obvious dimension along which the tasks vary is numbers used for each particular problem. For most diseases, the base rate is relatively low, and for most diagnostic tests in medicine, the hit rate (or sensitivity) is relatively high, and the false-alarm rate is relatively low. This was different in our four tasks (see **Table 1**), for which the base rates were—compared to most medical tasks—higher, the hit rates were lower, and the false-alarm rates were higher. Hence, it seems to be straightforward to explore the extent to which the base rate, the hit rate, and the false-alarm rate affect strategy use. When we started to do exactly this, it soon became evident that a larger database would be extremely useful, and so we also included the responses to the fifteen tasks of Gigerenzer and Hoffrage (1995) in the analysis. Moreover, we complemented the set of the three quantitative task variables with three qualitative dimensions—norm deviation, stakes, and main focus—and subsequently used these task characteristics to account for the variance of participants' responses and strategy use. This investigation, which can be considered as an example of an ecological analysis of Bayesian inferences, goes way beyond the scope of the present paper, and hence we report the results elsewhere (Hafenbrädl and Hoffrage, in review).

Differences between Undergraduates and Executives

We do not know why undergraduates outperformed the executives when information was presented in terms of natural frequencies, whereas executives outperformed the undergraduates when information was presented in terms of percentages. Formulating this finding as an interaction, though, may help to find a possible explanation. While representation format played a larger role for undergraduates, executives were relatively immune against this manipulation (**Figure 4**). In fact, within the sample of executives the effect of representation format (differences of 9.6 percentage points in favor of natural frequencies) did not reach significance ($p = 0.095$, which may, of course and as always for non-significant differences, be an issue of statistical power). There might be two ways to arrive at a response: arithmetic calculation and intuitive estimation. Maybe executives had a more intuitive approach, possibly based on their experience with similar problems in the world of business (Klein, 2002). If such experience is used, then representation format might indeed play less of a role. In contrast, undergraduates lack such experience and are hence more likely to approach the tasks with logic, reasoning, and arithmetic. The fact that natural frequencies facilitate the computation (Gigerenzer and Hoffrage, 1995) may hence account for the fact that undergraduates benefitted quite a lot from this representation—more than the executives did. But we must admit that this consideration is highly speculative and we should add that we did not find a similar pattern when comparing medical students (Hoffrage et al., 2000) to experienced physicians (Hoffrage and Gigerenzer, 1998).

Order Effects: Time Pressure and Training

More participants came up with the Bayesian response for tasks on the first page compared to tasks on the second page (difference of 2.5 percentage points). There are two main explanations for order effects: time pressure and training. If time pressure played a role, then we should expect that performance declines. In fact, out of those participants who provided an estimate to only one task, there were 28 who did so for the task on the first page, and 12 who did so only for the task on the second page (corresponding to a difference of 4.8 percentage points). In contrast, if training effects played a role, then we should expect that performance will increase. To the extent that the observed effect (2.5% better performance on first page) can be conceived as a result of a simple linear combination of the two possible components, time pressure and training, the effect of training is probably larger than the 2.5% that we observed, simply because this difference of 2.5% might have been overshadowed by the effect of time pressure. But we hasten to add that the observed difference was miniscule and of minor importance from a theoretical and practical point of view, and that our data does not allow us to assess the two possible contributing effects independently.

Teaching Bayesian Inferences

Another remarkable finding of the present study is that 69% of the undergraduates whom we asked said that they had heard of Bayes' rule, 68% said that they had been taught about it, 40% said they knew when it is applicable, 20% were actually able to correctly specify this, 32% said they were able to formulate Bayes' rule, and only 12% could actually do so (and even this number must probably be corrected downwards, see result section). These numbers suggest that one should not be too optimistic that teaching Bayes' rule leads to sustainable knowledge, retrievable from long-term memory. As one of the physicians studied in Hoffrage and Gigerenzer (1998) remarked to the experimenter after having filled out the questionnaire: "We have learned a formula at university, but I have forgotten it." Moreover, none of the variables concerning prior knowledge of Bayes' rule had a significant effect on Bayesian performance in our task. This lack of relationship could well be due to lack of statistical power, but if

91 participants are not enough to establish any relationship then such an effect, if existing at all, may be too small to be of practical importance.

Why is it that prior exposure to Bayes' rule seems to make almost no difference? We suspect that difficulties to remember Bayes' rule and to benefit from instructions may be related to how it is taught. In fact, inspecting an informal sample of textbooks on business statistics (Lawrence and Pasternack, 2002; Anderson et al., 2010; Newbold et al., 2010; Taylor, 2010) revealed the same picture as for the medical field: Bayes' rule is taught almost exclusively using probabilities. We agree that such textbooks must ensure that a student will, at the end of the lessons, be able to handle probability information, but we disagree that the best way to get there is to teach how to insert probabilities into Bayes' rule. Instead, we propose that students should be taught how to convert probabilities into natural frequencies. Sedlmeier and Gigerenzer (2001) have shown that a computerized implementation of such training is by far more effective compared to traditional rule training: the proportion of accurate answers doubled when participants had learned to represent probabilities as natural frequencies, as opposed to inserting them into Bayes' rule. Kurzenhäuser and Hoffrage (2002) obtained similar results in a classroom setting with medical students and diagnostic tasks from human genetics. Note that in both studies, the success of the two treatments—in one, students had been taught how to plug in probabilities into Bayes' rule, and in the other, they had been taught how to convert probability information into natural frequencies and derive the solution from there—has been evaluated by giving participants tasks with information presented in terms of probabilities. Taken together, the findings presented in this paper—supported by Sedlmeier and Gigerenzer's (2001) and Kurzenhäuser and Hoffrage's (2002) evaluation of tutorial programs—suggest that textbooks should no longer teach how to plug probabilities or percentages into a formula but rather instruct how to represent information in terms of natural frequencies to achieve a more sustainable mastery of Bayesian inference tasks.

Updating beliefs is a vital task, also in the domain of management. Representing information in terms of natural frequencies reduces computational complexity, improves understanding and boosts Bayesian performance. The posterior probability that managers can also benefit from natural frequencies, given the data of the present study, has definitely increased compared to the prior.

Acknowledgments

We would like to thank Justin Olds and the reviewers for helpful comments on a previous version, and Matthieu Legeret for assisting us with coding the data. This work was supported by grant 100014_140503 from the Swiss National Science Foundation.

References

Akl, E. A., Oxman, A. D., Herrin, J., Vist, G. E., Terrenato, I., and Sperati, F. (2011). Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database Syst. Rev.* 4, CD006776. doi: 10.1002/14651858.CD006776.pub2

Anderson, D. R., Sweeney, D. J., Williams, T. A., Freeman, J., and Shoesmith, E. (2010). *Statistics for Business and Economics*, 2nd Edn. Hampshire: Cengage Learning EMEA.

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

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychon. Bull. Rev.* 15, 284–289. doi: 10.3758/PBR.15.2.284

Brase, G. L., and Hill, W. T. (2015). Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Front. Psychol.* 6:340. doi: 10.3389/fpsyg.2015.00340

Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine: problems and opportunities," in *Judgment Under Uncertainty: Heuristics and Biases*, eds D. Kahneman, P. Slovic, and A. Tversky (Cambridge: Cambridge University Press), 249–267.

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

Gigerenzer, G. (1996). On narrow norms and vague heuristics: a reply to Kahneman and Tversky (1996). *Psychol. Rev.* 103, 592–596. doi: 10.1037/0033-295X.103.3.592

Gigerenzer, G. (2002). *Calculated Risks: How to Know When Numbers Deceive You*. New York: Simon & Schuster.

Gigerenzer, G. (2014). *Risk Savvy: How to Make Good Decisions*. New York: Viking Adult.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychol. Sci. Public Interest* 8, 53–96. doi: 10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G., and Gray, J. A. M. (eds). (2011). *Better Doctors, Better Patients, Better Decisions: Envisioning Healthcare in 2020*. Cambridge: MIT Press.

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., and Hoffrage, U. (2007). The role of representation in Bayesian reasoning: correcting common misconceptions. *Behav. Brain Sci.* 30, 264–267. doi: 10.1017/S0140525X07001756

Gigerenzer, G., and Murray, D. J. (1987). *Cognition as Intuitive Statistics*. Hillsdale, NJ: Erlbaum.

Groebner, D. F., Shannon, P. W., Philipp, C. F., and Smith, K. D. (2007). *Business Statistics: A Decision Making Approach*, 7th Edn. Upper Saddle River, NJ: Prentice Hall.

Hill, W. T., and Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *Q. J. Exp. Psychol.* 65, 2343–2368. doi: 10.1080/17470218.2012.687004

Hoffrage, U., and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Acad. Med.* 73, 538–540. doi: 10.1097/00001888-199805000-00024

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition* 84, 343–352. doi: 10.1016/S0010-0277(02)00050-1

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science* 290, 2261–2262. doi: 10.1126/science.290.5500.2261

Kahneman, D., Slovic, P., and Tversky, A. (eds). (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Klein, G. (2002). *The Power of Intuition: How to Use Your Gut Feelings to Make Better Decisions at Work*. New York, NY: Random House.

Kurzenhäuser, S., and Hoffrage, U. (2002). Teaching Bayesian reasoning: an evaluation of a classroom tutorial for medical students. *Med. Teach.* 24, 516–521. doi: 10.1080/0142159021000012540

Lawrence, J. A., and Pasternack, B. A. (2002). *Applied Management Science: Modeling, Spreadsheet Analysis, and Communication for Decision Making*, 2nd Edn. New Jersey: John Wiley & Sons, Inc.

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Front. Psychol.* 6:387. doi: 10.3389/fpsyg.2015.00387

Newbold, P., Carlson, W. L., and Thorne, B. (2010). *Statistics for Business and Economics*, 7th Edn. Upper Saddle River, NJ: Prentice Hall.

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Taylor, B. W. III. (2010). *Introduction to Management Science*, 10th Edn. Upper Saddle River, NJ: Prentice Hall.

Woloshin, S., and Schwartz, L. M. (2011). Communicating data about the benefits and harms of treatment: a randomized trial. *Ann. Intern. Med.*, 155, 87–96. doi: 10.7326/0003-4819-155-2-201107190-00004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix

Sentences or fragments of sentences that appear in parentheses and start with "P:" were only used in the percentage version, and those starting with "F:" were only used in the natural frequency version.

IRS Audit

This year experts project that (P: 20% of all) (F: 200 out of 1000) taxpayers will file an incorrect tax return. To identify such incorrect returns, the Internal Revenue Service (IRS) has been implemented. Unfortunately, this service is not perfect. IRS auditors detect an error for (P: 30% of the) (F: only 60 of those 200) tax returns that are incorrect, and it will indicate an error in (P: 10% of the) (F: 80 of these 800) tax returns that are correct.

The IRS has just notified a taxpayer there is an error in his return.

(P: What is the probability that the return actually has an error? _____ %)

(F: How many of the tax payers who have been notified by the IRS that there is an error in their return, do actually have an error? _____ of _____)

Techtronics Equipment Corporation

The Techtronics Equipment Corporation has developed a new electronic device that it would like to sell to the US Military for use in fighter aircraft. The sales manager knows that the military has placed an order (P: in 60% of) (F: in 60 of 100) similar cases. After making an initial sales presentation, military officials will often ask for a second presentation to other military decision makers. Historically, (P: 70% of successful companies are asked to make a second presentation, whereas only 50% of unsuccessful companies are asked back a second time.) (F: in 42 of the 60 successful cases, the companies were asked to make a second presentation, whereas for the unsuccessful cases, the companies were asked back a second time in only 20 of the 40 cases.)

Suppose Techtronics Equipment has just been asked to make a second presentation and so the sales manager wonders:

(P: What is the probability that the company will make the sale? _____ %)

(F: In how many of the cases in which a company has been called back, did this company receive an order? _____ of _____)

Varden Soap Company

The Varden Soap Company has two production facilities, one in Ohio and one in Virginia. The company makes the same type of soap at both facilities. (P: The Ohio plant makes 60% of the company's total soap output, and the Virginia plant 40%.) (F: Imagine 1000 containers of soap. The Ohio plant produces 600 of these containers, and the Virginia plant produces 400.) All soap from the two facilities is sent to a central warehouse, where it is intermingled. After extensive study, the quality assurance manager has determined that (P: 5% of the soap produced in Ohio and 10% of the soap in Virginia is) (F: 30 of the 600 soap containers produced in Ohio and 40 of the 400 soap containers produced in Virginia are) unusable due to quality problems. When the company sells a defective product, it incurs not only the cost of replacing the item but also the loss of goodwill. The vice president for production would like to allocate these costs fairly between the two plants.

(P: To do so, he wants to know, for instance: What is the probability that a soap was produced in Ohio given that it is defective? _____ %)

(F: To do so, he wonders, for instance: How many of the container with defective soap where produced in Ohio? _____ of _____)

frontiers in Psychology

ORIGINAL RESEARCH
published: 16 July 2015
doi: 10.3389/fpsyg.2015.00932

Visual aids improve diagnostic inferences and metacognitive judgment calibration

Rocio Garcia-Retamero [1,2,3], Edward T. Cokely [4,2,3] and Ulrich Hoffrage [5]*

[1] Department of Experimental Psychology, Facultad de Psicología, University of Granada, Granada, Spain, [2] Department of Cognitive and Learning Sciences, Michigan Technological University, Houghton, MI, USA, [3] Max Planck Institute for Human Development, Berlin, Germany, [4] National Institute for Risk and Resilience, University of Oklahoma, Norman, OK, USA, [5] Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland

Visual aids can improve comprehension of risks associated with medical treatments, screenings, and lifestyles. Do visual aids also help decision makers accurately assess their risk comprehension? That is, do visual aids help them become well calibrated? To address these questions, we investigated the benefits of visual aids displaying numerical information and measured accuracy of self-assessment of diagnostic inferences (i.e., metacognitive judgment calibration) controlling for individual differences in numeracy. Participants included 108 patients who made diagnostic inferences about three medical tests on the basis of information about the sensitivity and false-positive rate of the tests and disease prevalence. Half of the patients received the information in numbers without a visual aid, while the other half received numbers along with a grid representing the numerical information. In the numerical condition, many patients–especially those with low numeracy–misinterpreted the predictive value of the tests and profoundly overestimated the accuracy of their inferences. Metacognitive judgment calibration mediated the relationship between numeracy and accuracy of diagnostic inferences. In contrast, in the visual aid condition, patients at all levels of numeracy showed high-levels of inferential accuracy and metacognitive judgment calibration. Results indicate that accurate metacognitive assessment may explain the beneficial effects of visual aids and numeracy–a result that accords with theory suggesting that metacognition is an essential part of risk literacy. We conclude that well-designed risk communications can inform patients about health-relevant numerical information while helping them assess the quality of their own risk comprehension.

Keywords: visual aids, Bayesian reasoning, natural frequencies, numeracy, risk literacy, medical decision making, diagnostic inferences

OPEN ACCESS

Edited by:
David R. Mandel, Toronto Research Centre, Canada

Reviewed by:
Gary L. Brase, Kansas State University, USA
Nathan Dieckmann, Oregon Health & Science University, USA

****Correspondence:***
Rocio Garcia-Retamero, Department of Experimental Psychology, Facultad de Psicología, University of Granada, Campus Universitario de Cartuja s/n, 18071 Granada, Spain
rretamer@ugr.es

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *03 February 2015*
Accepted: *22 June 2015*
Published: *16 July 2015*

Citation:
Garcia-Retamero R, Cokely ET and Hoffrage U (2015) Visual aids improve diagnostic inferences and metacognitive judgment calibration. Front. Psychol. 6:932. doi: 10.3389/fpsyg.2015.00932

Introduction

Visual aids are graphical representations of numerical expressions of probability. They include, among others, icon arrays, bar and line charts, and grids (Paling, 2003; Spiegelhalter et al., 2011). Visual aids provide an effective means of risk communication when they are *transparent* (Garcia-Retamero and Cokely, 2013)—that is, when their elements are well defined and they accurately and clearly represent the relevant risk information by making part-to-whole relationships in the data

visually available (Gillan et al., 1998; Ancker et al., 2006; Reyna and Brainerd, 2008; Fischhoff et al., 2012; Trevena et al., 2012).

Transparent visual aids improve comprehension of risks associated with different lifestyles, screenings, and medical treatments, and they promote consideration of beneficial treatments despite side-effects (Feldman-Stewart et al., 2000; Paling, 2003; Waters et al., 2007; Zikmund-Fisher et al., 2008a; Zikmund-Fisher, 2015). Transparent visual aids also increase appropriate risk-avoidance behaviors, they promote healthy behaviors (Garcia-Retamero and Cokely, 2011, 2014a), they reduce errors and biases induced by anecdotal narratives and framed messages (Fagerlin et al., 2005; Schirillo and Stone, 2005; Garcia-Retamero and Galesic, 2009, 2010a; Cox et al., 2010; Garcia-Retamero et al., 2010) and they aid comprehension of complex concepts such as incremental risk (Zikmund-Fisher et al., 2008b). Risk information presented visually is also judged as easier to understand and recall than the same information presented numerically (Feldman-Stewart et al., 2007; Goodyear-Smith et al., 2008; Gaissmaier et al., 2012; Zikmund-Fisher et al., 2014; Okan et al., 2015).

However, not all visual aids are equally effective for all tasks (see Garcia-Retamero and Cokely, 2013, for a review). For instance, bar graphs are useful for comparing data points (Lipkus and Hollands, 1999; Lipkus, 2007; Fischhoff et al., 2012); line graphs are helpful for depicting trends over time; magnifier risk scales (including magnifying lenses) are useful for depicting small numbers (Ancker et al., 2006); icon arrays can be helpful for communicating treatment risk reduction and risk of side effects (Feldman-Stewart et al., 2000; Garcia-Retamero and Galesic, 2009, 2010b; Ancker et al., 2011; Okan et al., 2012); logic trees can be useful for visually depicting argument structure (Mandel, 2014); and grids can help depict large numbers when communicating the predictive value of medical tests (Garcia-Retamero and Hoffrage, 2013).

Grids displaying numerical information graphically have been found to boost the accuracy of perceptions of health-related benefits and risks beyond the effect of other transparent information formats. To illustrate, doctors and patients often have difficulties inferring the predictive value of a medical test from information about the sensitivity and false-positive rate of the test and the prevalence of the disease. In an influential study on how doctors process information about the results of mammography, Eddy (1982) gave 100 doctors the following information: "The probability that a woman has breast cancer is 1%. When a woman has breast cancer, it is not sure that she will have a positive result on the mammography: she has an 80% probability of having a positive result on the mammography. When a woman does not have breast cancer, it is still possible that she will have a positive result on the mammography: she has a 10% probability of having a positive result on the mammography."

After having read this information, doctors were required to estimate the probability that a woman with a positive mammography actually has breast cancer. Eddy (1982) reported that 95 of 100 doctors estimated this probability to be about 80% (see Gigerenzer, 2013; Ellis et al., 2014, for similar results in patients). If one inserts the numbers presented above into a Bayes' theorem, however, one gets a value of 8%, which is one order of magnitude smaller.

Gigerenzer and Hoffrage (1995, 1999) showed that communicating information about medical tests in natural frequencies as compared to probabilities improves diagnostic inferences (see also Sedlmeier and Gigerenzer, 2001; Kurzenhäuser and Hoffrage, 2002; Mandel, 2015). Natural frequencies are final tallies in a set of objects or events randomly sampled from the natural environment (Hoffrage et al., 2000, 2002). For the mammography task the statistical information provided in terms of natural frequencies reads: "100 out of every 10,000 women have breast cancer. When a woman has breast cancer, it is not sure that she will have a positive result on the mammography: 80 of every 100 such women will have a positive result on the mammography. When a woman does not have breast cancer, it is still possible that she will have a positive result on the mammography: 990 out of every 9,900 such women will have a positive result on the mammography."

Even though the effect of numerical format (probabilities vs. natural frequencies) is substantial, performance in the natural frequency condition still leaves room for improvement. A study conducted by Garcia-Retamero and Hoffrage (2013) showed that grids displaying numerical information graphically improved diagnostic inferences in both doctors and their patients beyond the effect of natural frequencies (see also Brase, 2014, for similar results in young adults). The authors showed that these grids not only increased objective accuracy but also increased perceived usefulness of information and decreased perceived task difficulty. The aim of the current research was to extend this literature by investigating whether visual aids also help decision makers accurately assess their risk comprehension (metacognitive judgment calibration). In particular, we followed the method used by Garcia-Retamero and Hoffrage (2013) and investigated whether grids graphically displaying information about the predictive value of medical tests improve self-assessment of diagnostic inferences in patients.

Previous research showed that people can be highly overconfident when assessing the accuracy of their own judgments (Griffin and Brenner, 2004). For example, Dunning et al. (2004) conducted a systematic review of the literature on the topic and concluded that people's self-views hold only a tenuous to modest relationship with their actual behavior and performance. On average, people say that they are "above average" in skill—a conclusion that defies statistical possibility for symmetric distributions of individuals (however, this conclusion is plausible if the mean and the median of a distribution are not identical; Gigerenzer et al., 2012). People also overestimate the likelihood that they will engage in desirable behaviors and achieve favorable outcomes, they furnish overly optimistic estimates of when they will complete future projects, and they reach judgments with too much confidence.

People tend to be highly overconfident at low levels of accuracy yet relatively well calibrated at higher levels of accuracy—a result that suggests the presence of an "unskilled and unaware effect" (Ehrlinger and Dunning, 2003; Ehrlinger et al., 2008). This result is consistent with research on individuals with low numeracy (i.e., the ability to accurately interpret numerical information about

risk; Ancker and Kaufman, 2007; Fagerlin et al., 2007; Reyna et al., 2009; Galesic and Garcia-Retamero, 2010; Cokely et al., 2012; Peters, 2012). This research shows that people with low numeracy are especially inaccurate when evaluating the accuracy of their own judgments, showing overconfidence (Ghazal et al., 2014), and are not able to use risk reduction information to adjust their estimates (Schwartz et al., 1997). Overconfidence mediates, at least in part, the effect of numeracy on judgment accuracy (Ghazal et al., 2014). Thus, people with low numeracy may struggle to grasp numerical concepts that are essential for understanding health-relevant information because they have difficulties assessing the accuracy of their own estimates.

Our hypothesis is that visual aids can improve both accuracy of diagnostic inferences and metacognitive judgment calibration (i.e., how well patients assess the accuracy of these inferences) (H_1). We also hypothesize that visual aids may be especially useful for patients with low numeracy (H_2). Visual aids can increase the likelihood that less numerate patients deliberate on the available risk information, elaborating more on the problem at hand and on their own understanding of the problem (Garcia-Retamero and Cokely, 2013, 2014b). Deliberation tends to be important for risk understanding because it promotes more thorough, complex, and durable information representations (Cokely and Kelley, 2009)—an important component of metacognitive judgment calibration (Thompson et al., 2011). By influencing encoding and representation, visual aids can increase metacognitive judgment calibration, reducing overconfidence. Improvements in metacognitive processes can, in turn, improve the accuracy of inferences (H_3).

Materials and Methods

Participants

Participants included 108 patients recruited from four hospitals in the cities of Jaén and Granada (Spain) during treatment consultation. To be eligible for recruitment, patients had to have no previous formal medical training. If they agreed to participate, they were provided with an introductory letter describing the purpose of the study and their questions were answered. Eighty four percent of the patients who had been approached ($n = 128$) agreed to participate in the study. Those who refused mentioned one or more of the following reasons: respondent burden, lack of interest in research, and/or busy schedules. Patients had an average age of 52 years (range 19–76), and 78% were females. Most of the patients (86%) had a high school degree or less, and only 14% had a university education before participating in the study. Twenty-three percent of the patients had a chronic condition (e.g., allergies or diabetes). Patients received €20 for participating in the study and were assigned randomly to one of two groups. Male and female patients were evenly distributed in the groups. The Ethics Committee of the University of Granada approved the methodology, and all patients consented to participation through a consent form at the beginning of the study.

Materials and Procedure

Patients completed a two-part paper-and-pencil questionnaire. In the first part, they were presented with three tasks involving different diagnostic inferences: inferring breast cancer from a positive mammogram, colon cancer from a positive hemoccult test, and insulin-dependent diabetes from a genetic test. The order of the three tasks was randomized, independently for each patient. Wording and length of the tasks were comparable to the variant of the breast cancer task that we provided in the introduction of the current article. The information about the sensitivity and false-positive rate of the tests and prevalence of the diseases was taken from published studies (Hoffrage and Gigerenzer, 1998; Garcia-Retamero and Hoffrage, 2013) and was reported in natural frequencies (see **Table 1**). There were no time constraints, but the questionnaire took approximately 15 min to complete.

TABLE 1 | Information about prevalence of the diseases, and sensitivity and false-positive rate of the tests.

Diagnostic task	Base rate	Sensitivity	False-positive rate	Positive predictive value
Breast cancer	100 of 10,000	80 of 100	990 of 9,900	80 of 1,070
Colon cancer	30 of 10,000	15 of 30	299 of 9,970	15 of 314
Diabetes	50 of 10,000	48 of 50	4,975 of 9,950	48 of 5,023

Note that the false-positive rate is the complement of the specificity.

Half of the patients received the information about the sensitivity and false-positive rate of the tests and prevalence of the diseases in numbers without a visual aid. The other half received numbers along with a grid representing the numerical information. **Figure 1** presents the grid that patients received in the mammography task. The visual display represented the number of women who obtained a positive mammogram, the number of women who have breast cancer, and the overall number of women at risk. Women were depicted as squares as previous research has found no differences in effects of arrays with faces compared to more abstract symbols such as squares or circles (Stone et al., 2003; Gaissmaier et al., 2012).

After having received the information about the sensitivity and false-positive rate of the test and the base rate of the disease for a given task, patients made a diagnostic inference. In the breast cancer task, patients were told: "Imagine a representative sample of women who got a positive result on the mammography. Give your best guess: how many of these women do you expect to have breast cancer?" Patients were asked to provide two numbers such as X out of Y (leaving it up to them which denominator to use). After making the three diagnostic inferences, patients estimated accuracy of their diagnostic inferences. In particular, they estimated the number of correct inferences that they thought they had made on a scale ranging from 0 to 3. The second part of the questionnaire included a measure of numerical skills using twelve items taken from Schwartz et al. (1997) and Lipkus et al. (2001 ; see Cokely et al., 2012, for a review).

Design and Dependent Variables

We employed a mixed design with one independent variable manipulated experimentally between-groups: information format (numerical only vs. numerical and visual). In addition, we considered one independent variable that was not manipulated

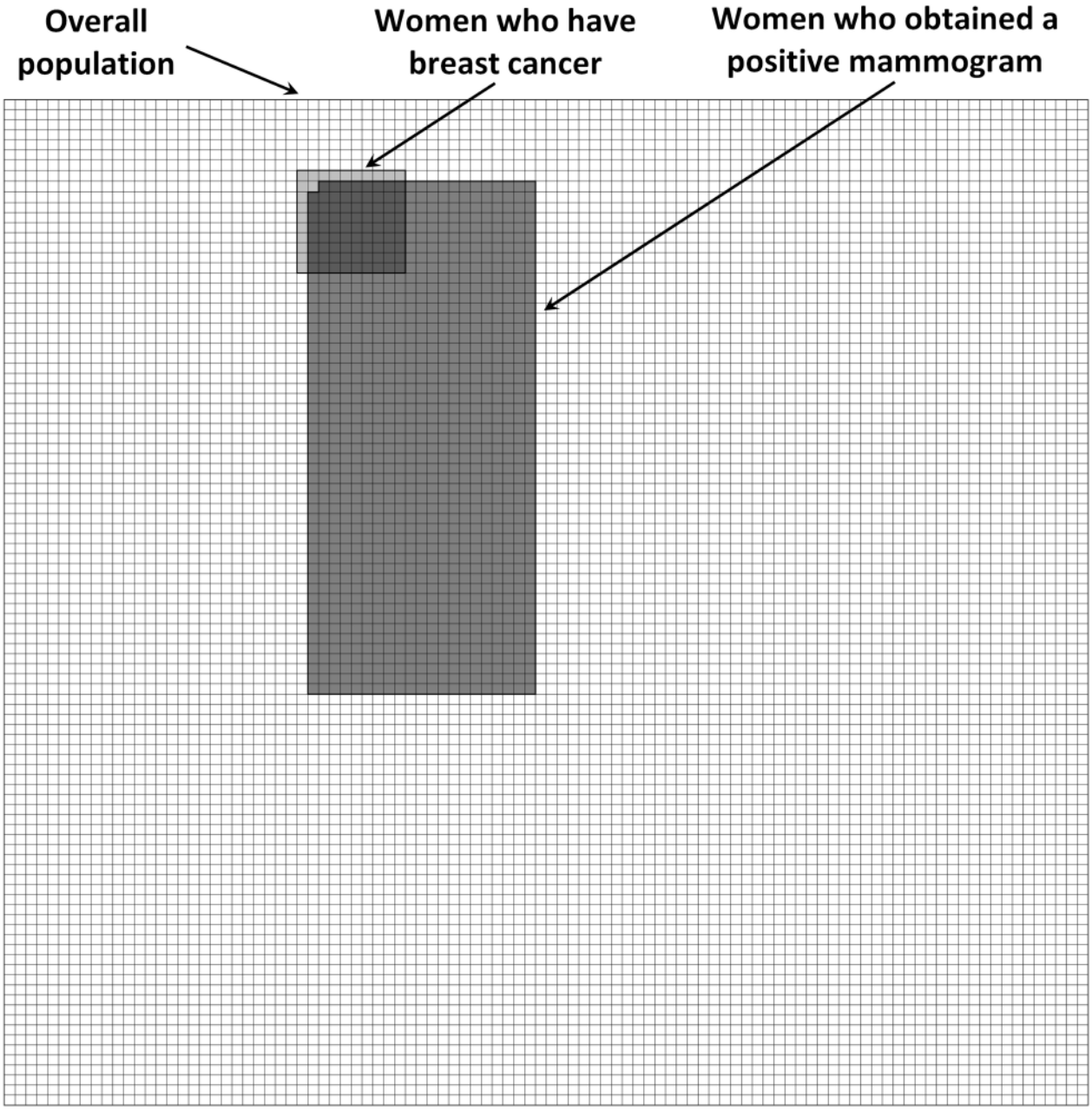

FIGURE 1 | Visual aid representing the overall number of women at risk, the number of women who have breast cancer, and the number of women who obtained a positive mammogram.

experimentally but measured, namely numeracy. We split patients into two groups according to the median of their numeracy scores. The low-numeracy group ($n = 52$) included patients with eight or fewer correct answers, while the high-numeracy group ($n = 56$) included those with nine or more correct answers (see Peters et al., 2006; Garcia-Retamero and Galesic, 2010b; and Garcia-Retamero and Cokely, 2014a, for a similar procedure).

Patients answered questions about the three tasks involving different diagnostic inferences. We used patients' answers to the questions to determine our three dependent variables. *Objective accuracy* was measured as the percentage of correct inferences in the three tasks. Following Gigerenzer and Hoffrage (1995; see also Hoffrage et al., 2000), a response was considered accurate if it matched the value specified in the last column of **Table 1** plus/minus one percentage point. A more liberal criterion than the one that we used in our analyses yielded similar findings to those reported in the results section. *Estimated accuracy* was measured as the estimated percentage of correct inferences in the three tasks. Finally, *metacognitive judgment calibration* was determined for each patient by computing the difference between estimated

accuracy and objective accuracy (see Ghazal et al., 2014, for a similar method).

Analyses

First, we conducted analyses of variance (ANOVAs) to assess the effect of information format and numeracy on objective accuracy, estimated accuracy, and metacognitive judgment calibration (H_1 and H_2). Second, we assessed whether metacognitive judgment calibration explains the effect of information format and numeracy on objective accuracy (H_3). In particular, we conducted an analysis of covariance (ANCOVA) to assess the effect of information format and numeracy on objective accuracy after controlling for metacognitive judgment calibration. We also conducted mediational analyses to assess whether the effect of information format and numeracy on objective accuracy was mediated by metacognitive judgment calibration.

Finally, to find additional support of our hypothesis (H_3) and address an alternative explanation of our results, we investigated whether objective accuracy explains the effect of information format and numeracy on metacognitive judgment calibration. In particular, we conducted an ANCOVA to assess the effect of information format and numeracy on metacognitive judgment calibration after controlling for objective accuracy. In addition, we conducted mediational analyses to assess whether the effect of information format and numeracy on metacognitive judgment calibration was mediated by objective accuracy. As this alternative model seems plausible, we compared the size of its indirect effect (i.e., the amount of mediation) with that of the model with metacognitive judgment calibration as a mediator. Numeracy was included as a dichotomous variable in the ANOVAs and ANCOVAs and as a continuous variable in the mediation analyses. We found consistent results in these analyses (for a similar method, see Peters et al., 2006; Garcia-Retamero and Galesic, 2009, 2010b; and Garcia-Retamero and Cokely, 2014a).

Results

How did patients perform in the diagnostic inference tasks? And how did they think they had performed in the tasks? The percentage of patients who answered correctly three, two, one, and zero tasks was 24, 19, 18, and 39% respectively. In contrast, 50, 25, 19, and 6% of the patients estimated that they had made three, two, one, and zero correct diagnostic inferences, respectively. Only 34% of the patients were accurate when assessing the accuracy of their inferences (i.e., they were well calibrated); 38% overestimated accuracy in one task (33%); 18% overestimated accuracy in two tasks (67%); 6% overestimated accuracy in three tasks (100%); and 4% underestimated accuracy. Patients who achieved higher levels of accuracy were well calibrated, whereas patient with low levels of accuracy were highly overconfident (see **Figure 2**).

Do visual aids and numeracy affect objective accuracy? Are visual aids especially useful for patients with low numeracy? Patients made more accurate inferences when the information was presented both numerically and visually (55% correct inferences) as compared to numerically only (32%) (H_1). In addition, patients with high numeracy were more accurate (51% correct inferences) as compared to low-numerate patients (35%). Finally, grids displaying numerical information were particularly useful additions for patients with low numeracy (see **Figure 3**). In contrast, there was only a minor increase in accuracy in patients with high numeracy when they received the additional visual display (H_2). In line with these results, the ANOVA with information format and numeracy as between-subjects factors and objective accuracy across the three tasks as the dependent variable revealed a main effect of information format, $F_{1,104} = 10.77$, $p = 0.001$, $\eta_p^2 = 0.09$, and numeracy, $F_{1,104} = 5.79$, $p = 0.02$, $\eta_p^2 = 0.05$. The interaction between information format and numeracy was also significant, $F_{1,104} = 3.82$, $p = 0.05$, $\eta_p^2 = 0.04$.

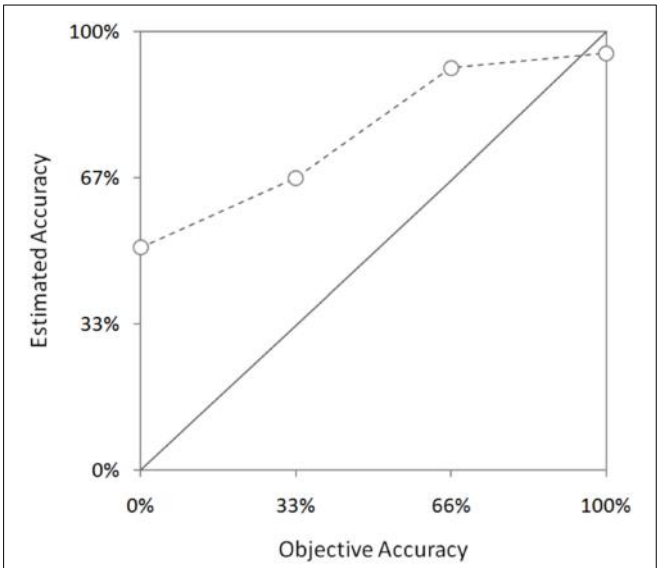

FIGURE 2 | Estimated accuracy by objective accuracy. Error bars indicate one standard error of the mean.

Do visual aids and numeracy affect estimated accuracy and metacognitive judgment calibration? Are visual aids especially useful for patients with low numeracy? Estimates of accuracy were not influenced by information format or numeracy. On average, patients estimated that 72% of their inferences were correct (see **Figure 3**). In contrast, both information format and numeracy had an effect on accuracy of estimates (i.e., metacognitive judgment calibration) (H_1). Grids displaying numerical information improved metacognitive judgment calibration in patients with low numeracy. These patients more accurately estimated the accuracy of their own inferences when they received the visual aid. However, the beneficial effect of the visual aid could not be observed in patients with high numeracy (H_2). These patients were relatively well calibrated regardless of information format. In line with these results, the ANOVA with information format and numeracy as between-subjects factors and estimated accuracy of diagnostic inferences across the three tasks as a dependent variable did not reveal any significant results ($F < 1$). In contrast, the ANOVA with information format and numeracy as between-subjects factors and metacognitive judgment calibration as a dependent variable revealed a main effect of information format, $F_{1,104} = 14.62$, $p = 0.001$, $\eta_p^2 = 0.12$, and numeracy, $F_{1,104} = 5.28$,

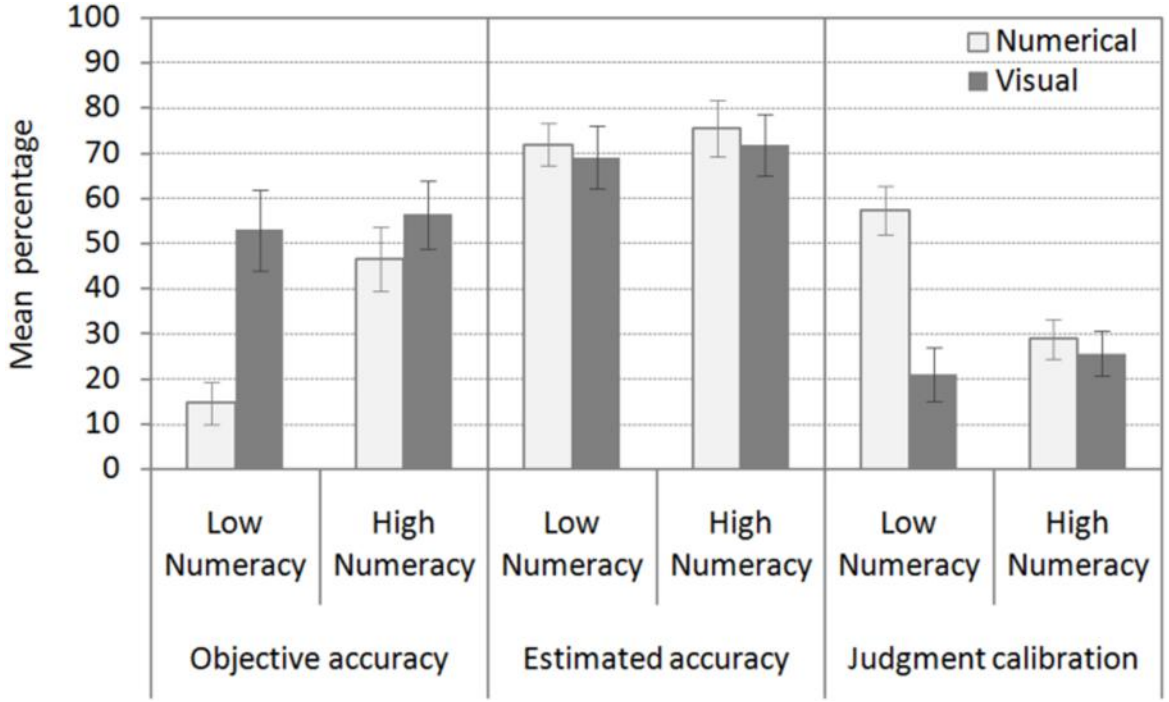

FIGURE 3 | Objective accuracy, estimated accuracy, and metacognitive judgment calibration across the three diagnostic tasks by information format and numeracy. Error bars indicate one standard error of the mean.

$p = 0.02$, $\eta_p^2 = 0.05$, and an interaction between information format and numeracy, $F_{1,104} = 10.22$, $p = 0.002$, $\eta_p^2 = 0.09$.

Does metacognitive judgment calibration explain the effect of information format and numeracy on objective accuracy? Visual aids do not improve objective accuracy in patients with low numeracy when metacognitive judgment calibration has been controlled for statistically (see **Figure 4**). In line with these results, the ANCOVA with information format and numeracy as between-subjects factors, objective accuracy across the three tasks as the dependent variable, and metacognitive judgment calibration as a covariate only revealed a main effect of metacognitive judgment calibration, $F_{1,103} = 37.25$, $p = 0.001$, $\eta_p^2 = 0.27$. The main effect of information format and numeracy and the interaction between information format and numeracy was no longer significant ($F < 1$).

To ensure comparability with results in the ANOVA and ANCOVA, in mediational analyses we first modeled objective accuracy when patients received information in numbers and then when they received an additional visual display representing the numerical information. In the numerical condition, regression analyses showed that numeracy influenced both metacognitive judgment calibration, $\beta = -0.56$, $t_{53} = -4.97$, $p = 0.001$, and objective accuracy, $\beta = 0.46$, $t_{53} = 3.82$, $p = 0.001$, whereby patients who were more numerate more accurately assessed the accuracy of their inferences (i.e., were better calibrated) and made more accurate inferences (see **Figure 5A**). In addition, metacognitive judgment calibration was related to objective accuracy, $\beta = -0.52$, $t_{52} = -4.04$, $p = 0.001$. Patients who more accurately assessed the accuracy of their inferences also made more accurate inferences. When metacognitive judgment calibration was included in the regression analyses, the effect of numeracy on objective accuracy was significantly reduced and was no longer significant, $\beta = 0.17$, $t_{52} = 1.30$, $p = 0.20$. The results of the Sobel test indicated that metacognitive judgment calibration mediates the relationship between numeracy and objective accuracy, $z = 3.135$, $p = 0.001$ [Effect = 0.30, 95% CI (0.27,0.33); AIC (Akaike Information Criterion) = 998.80]. When patients received the additional visual aid representing the numerical information, numeracy did not influence metacognitive judgment calibration, $\beta = 0.10$, $t_{51} = 0.70$, $p = 0.49$, or objective accuracy, $\beta = 0.14$, $t_{51} = 1.01$, $p = 0.32$ (see **Figure 5B**). As expected, metacognitive judgment calibration was again related to objective accuracy, $\beta = -0.53$, $t_{50} = -4.48$, $p = 0.001$.

Does objective accuracy explain the effect of information format and numeracy on metacognitive judgment calibration? Visual aids improve metacognitive judgment calibration in patients with low numeracy after objective accuracy has been controlled for statistically (see **Figure 4**). The ANCOVA with information format and numeracy as between-subjects factors, metacognitive judgment calibration across the three tasks as the dependent variable, and objective accuracy as a covariate revealed a main effect of objective accuracy, $F_{1,103} = 37.25$, $p = 0.001$, $\eta_p^2 = 0.27$, and information format, $F_{1,103} = 5.56$, $p = 0.020$, $\eta_p^2 = 0.05$, and an interaction between information format and numeracy, $F_{1,103} = 6.24$, $p = 0.014$, $\eta_p^2 = 0.06$.

As expected, in the numerical condition, regression analyses showed that objective accuracy was related to metacognitive judgment calibration, $\beta = -0.46$, $t_{52} = -4.04$, $p = 0.001$ (see

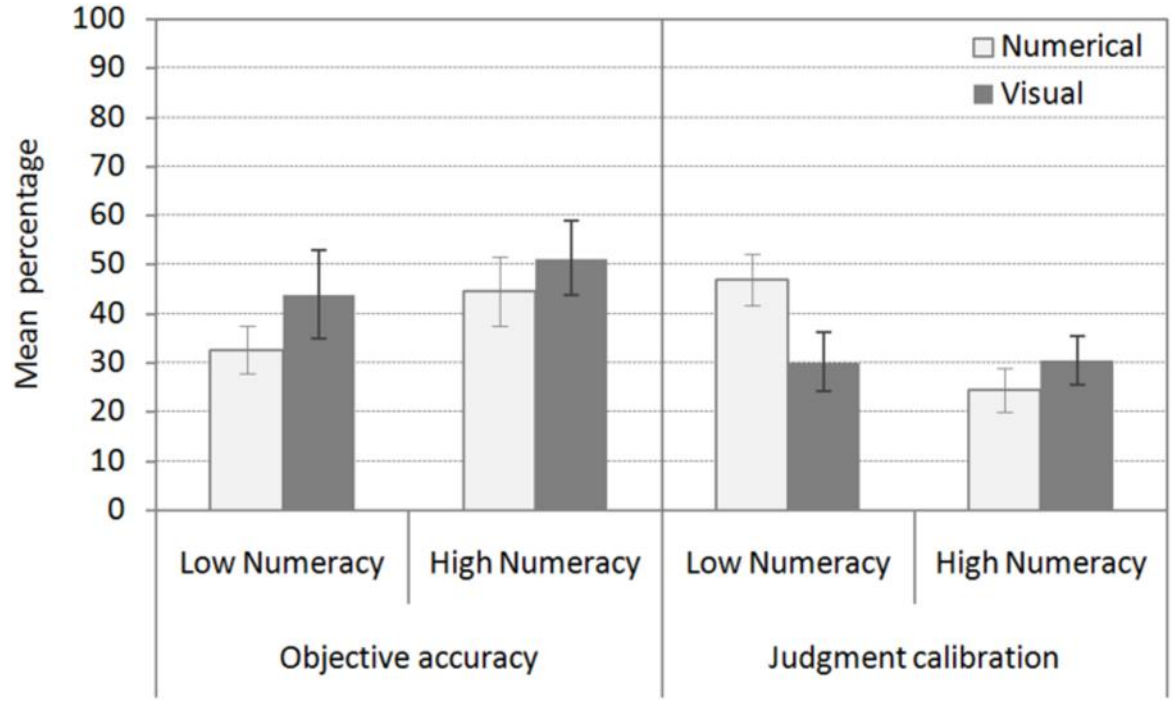

FIGURE 4 | Objective accuracy across the three diagnostic tasks by information format and numeracy after controlling for the effect of metacognitive judgment calibration. Metacognitive judgment calibration across the three diagnostic tasks by information format and numeracy after controlling for the effect of objective accuracy. Error bars indicate one standard error of the mean.

Figure 5C). Patients who made more accurate inferences also more accurately assessed the accuracy of these inferences. When objective accuracy was included in the regression analyses, the effect of numeracy on metacognitive judgment calibration was reduced but it was still significant, $\beta = -0.35$, $t_{52} = -3.12$, $p = 0.003$. The results of the Sobel test indicated that objective accuracy mediates the relationship between numeracy and metacognitive judgment calibration, $z = -2.78$, $p = 0.003$. However, the size of the indirect effect [Effect $= -0.21$, 95% CI $(-0.24, -0.18)$] was smaller and AIC (AIC $= 1057.30$) was larger to that of the previous model. These results suggest that the model including objective accuracy as a mediator is a worse model than the model including metacognitive judgment calibration as a mediator.

In line with previous results, when patients received the additional visual aid representing the numerical information, objective accuracy was related to metacognitive judgment calibration, $\beta = -0.54$, $t_{50} = -4.48$, $p = 0.001$ (see **Figure 5D**). In sum, results in ANCOVAs and mediational analyses suggest that metacognitive judgment calibration mediates the effect of numeracy on objective accuracy (H_3) and not the other way around. Thus these analyses suggest that, in the numerical condition, highly numerate patients make more accurate inferences than patients with low numeracy because they more accurately evaluate the accuracy of their own inferences. In contrast, in the visual condition, patients at all levels of numeracy showed similar high-levels of metacognitive judgment calibration and, in turn, high-levels of inferential accuracy.

Discussion

We investigated patients' diagnostic inferences about the predictive value of medical tests from information about the sensitivity and false-positive rate of the tests and the prevalence of several diseases. Our results showed that many patients—especially those with low numeracy—made incorrect inferences about the predictive value of the tests and dramatically overestimated the accuracy of these inferences. High overestimates at low levels of accuracy become more calibrated at higher levels of accuracy—a result that suggests the presence of an "unskilled and unaware effect" (see also Ehrlinger and Dunning, 2003; Ehrlinger et al., 2008; Ghazal et al., 2014).

Our results are compatible with previous evidence on the role of numeracy in understanding health-relevant risk communications and medical decision making (Fagerlin et al., 2007; Apter et al., 2008; Reyna et al., 2009; Peters, 2012; Garcia-Retamero and Galesic, 2013; Johnson and Tubau, 2015). Patients with low levels of numeracy have more difficulties interpreting numerical risks of side effects (Gardner et al., 2011), and they are more susceptible to being influenced by the way the health information is framed in problems involving probabilities (Peters et al., 2006; Peters and Levin, 2008; Garcia-Retamero and Galesic, 2010a, 2011; Galesic and Garcia-Retamero, 2011a)—presumably because they are more influenced by non-numerical information (e.g., mood states; Peters et al., 2007; Petrova et al., 2014). Compared to patients with high numeracy, less-numerate patients also tend to overestimate their risk of suffering from several diseases (Davids

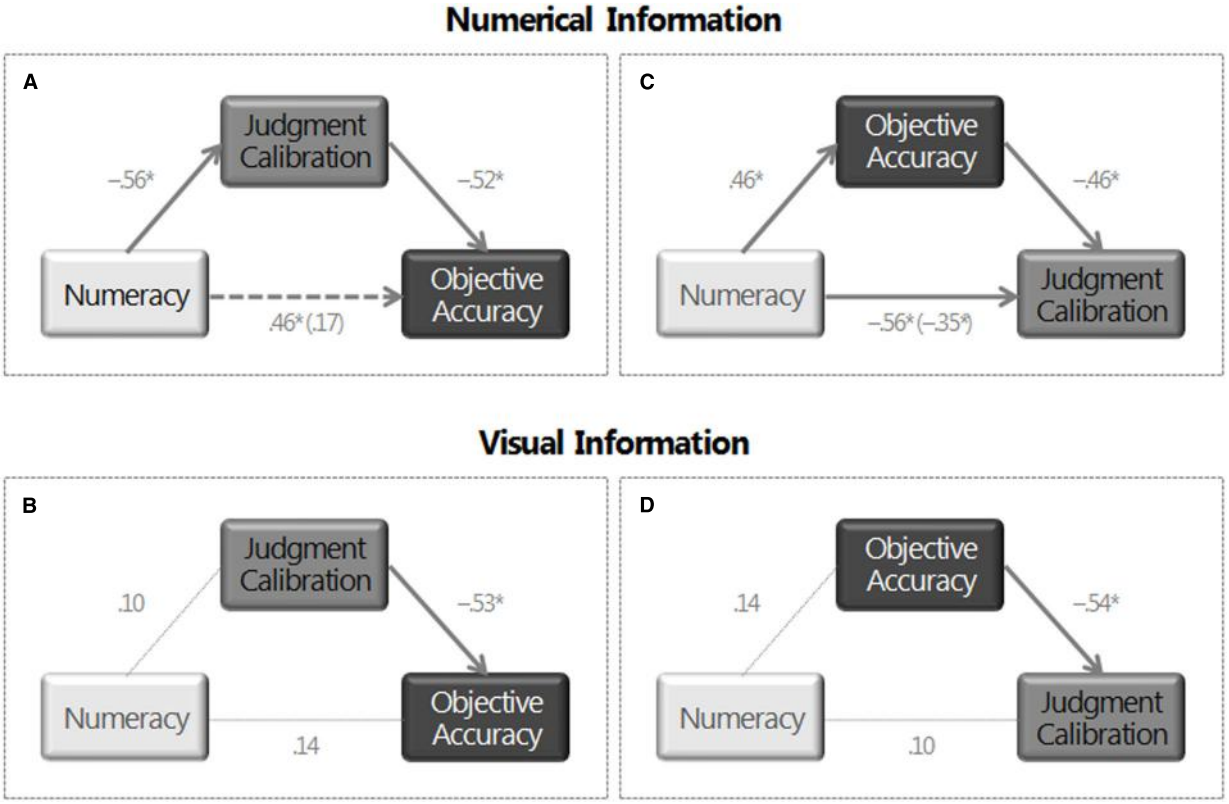

FIGURE 5 | Path analyses. Effect of numeracy on objective accuracy and the mediational effect of metacognitive judgment calibration when **(A)** patients received information only in numbers and when **(B)** they received an additional visual display representing the numerical information. Effect of numeracy on metacognitive judgment calibration and the mediational effect of objective accuracy when **(C)** patients received information only in numbers and when **(D)** they received an additional visual display representing the numerical information. Note: Standardized coefficients are shown. $^*p < 0.05$.

et al., 2004; Gurmankin et al., 2004), they are less able to use risk reduction information to adjust their risk estimates (e.g., screening data; Schwartz et al., 1997), they tend to overestimate benefits of uncertain treatments (Weinfurt et al., 2003; Garcia-Retamero and Galesic, 2010b), and they have more deficits in understanding the information necessary to follow dietary recommendations (Rothman et al., 2006). Compared to patients with high numeracy, less-numerate patients also tend to search for less information about their disease (Portnoy et al., 2010), and they often choose lower-quality health options (e.g., health insurance plans; Hibbard et al., 2007; Hanoch et al., 2010). As a consequence, they tend to suffer more comorbidity and take more prescribed drugs (Garcia-Retamero et al., 2015). Less-numerate doctors and patients also favor a paternalistic model of medical decision making, in which doctors are dominant and autonomous (Garcia-Retamero et al., 2014), and patients prefer not to participate and instead delegate decision making (Galesic and Garcia-Retamero, 2011b). This is troubling given that the paternalistic model of medical decision making is increasingly being questioned (Kaplan and Frosch, 2005).

Our research suggests a potential explanation of the link between numeracy and understanding of health-relevant quantitative information. Highly numerate patients might make more accurate inferences as compared to patients with low numeracy because they more accurately evaluate the accuracy of their own inferences (i.e., they show better metacognitive judgment calibration). Thus metacognitive judgment calibration might drive, at least in part, the numeracy-to-performance relationship. Previous research suggests that the link between numeracy and superior judgment and decision making might reflect differences in heuristic-based deliberation (e.g., deep elaborative processing; Cokely and Kelley, 2009; Cokely et al., 2012), affective numerical intuition (e.g., precise symbolic number mapping; Peters et al., 2006; Peters, 2012), and meaningful intuitive understanding (e.g., gist-based representation and reasoning; Reyna, 2004; Reyna et al., 2009; see Cokely et al., 2014, for a review). Our research extends this literature suggesting that there is also a tight link between numeracy, metacognition, and understanding of health-relevant numerical information (see Ghazal et al., 2014, for similar results in highly educated samples).

Our results are also compatible with a variety of studies indicating that judgment self-assessment can operate as a domain-general skill that correlates with—but that can also be seen as an independent predictor of—general abilities, personality traits, and cognitive performance (Stankov, 2000; Stankov and Lee, 2008; Schraw, 2010). Overall our results accord with metacognitive theory suggesting that metacognitive judgment

calibration tends to be useful because it is instrumental in self-regulation—i.e., the monitoring and control of cognition (Nelson, 1990; Metcalfe and Finn, 2008). Related studies of factors like "feeling of correctness" show that confidence-type judgments predict differences in information search and elaboration. In addition to predicting judgments about the correctness of one's answer, one's feeling of correctness tends to be related to "rethinking" times and the likelihood of changing one's initial answer during reasoning (Thompson et al., 2011). These studies suggest that factors related to how one uses and assesses judgment accuracy may often be essential components determining the extent to which one deliberates during judgment and decision making (Ghazal et al., 2014). For these and other reasons it seems likely that metacognition is an essential component of the ability to understand and make good decisions about risk (i.e., risk literacy; see www.RiskLiteracy.org).

Finally, our results can have important implications for medical practice as they suggest suitable ways to communicate quantitative medical data—especially to patients lacking numerical skills. Our research shows that visual aids improve both objective accuracy and metacognitive judgment calibration, especially in less numerate patients, eliminating differences between this group of patients and the more numerate group. In addition, our research suggests that visual aids increase objective accuracy by improving metacognitive judgment calibration. As we mentioned above, calibration can mediate the relationship between numeracy and superior performance. In the current research, however, this result only holds when patients received numerical information without a visual display. In contrast, metacognitive judgment calibration did not mediate the effect of numeracy on objective accuracy when patients received the additional visual aid representing the numerical information because numeracy was no longer as robustly related to accuracy of inferences. In the visual condition, both patients with low and high numeracy were often well calibrated and, in turn, often made accurate inferences. These results suggest that visual aids might improve risk understanding, at least in part, by improving metacognitive judgment calibration and reducing overestimates of accuracy.

It is also possible that the effect of visual aids on both judgment accuracy and metacognitive judgment calibration follow from the development of better cognitive representations, which, in turn, facilitate reasoning and metacognitive monitoring (see Cosmides and Tooby, 1996; Brase et al., 1998; Brase, 2009). For instance, more cues available in memory can be used to explore essential relationships or to recognize that one has some missing knowledge. This conclusion is compatible with previous research indicating that visual aids help less numerate people identify and infer essential aspects of the risk information (e.g., "gross-level information"; Feldman-Stewart et al., 2000; Zikmund-Fisher et al., 2010). Visual aids also increase the ability of less numerate people to recognize superordinate classes, making part-to-whole relations in the data visually available (Ancker et al., 2006; Reyna and Brainerd, 2008). Moreover, visual aids improve risk comprehension by increasing the likelihood that less numerate people deliberate on the available risk information (Garcia-Retamero and Cokely, 2013, 2014b). By influencing memory encoding and representation, visual aids can also give rise to enduring changes in attitudes and behavioral intentions, which in turn affect behavior and risky decision making (Garcia-Retamero and Cokely, 2011, 2014a, 2015). Thus visual aids can improve judgment and decision making and help promote healthy behavior by improving understanding of health-relevant numerical information, by improving assessments of the accuracy of inferences about this information, and by establishing enduring attitudes and fostering intentions to perform the behavior, which may further promote understanding and self-assessment.

As with any research, our study has some limitations and leaves open several questions for future research. For instance, objective accuracy and metacognitive judgment calibration were correlated as the former was included in the measurement of the latter. To the extent that judgment calibration cannot be defined independently of objective accuracy, these concepts are not independent. So any results in this area need to be benchmarked accordingly. Nevertheless, our analyses showed that information format and numeracy have a significant effect on metacognitive judgment calibration even after objective accuracy has been controlled for statistically.

It is important to mention that our conclusions are based on patients' diagnostic inferences and estimates when they received information about prevalence of several diseases, and the sensitivity and false-positive rate of the tests in natural frequencies (Hoffrage and Gigerenzer, 1998; Hoffrage et al., 2000, 2002). Future research could investigate these inferences and estimates when the information is reported in other numerical formats (e.g., probabilities). In addition, future research could also investigate whether these inferences and estimates affect behavioral intentions and actual behavior (e.g., whether patients indicate that they would take a medical test depending on the way the information about the test is communicated and if expressed interest exceeds actual uptake). Our sample of patients was older and less educated than the general population in Spain and other countries. Future research could also examine whether visual aids confer similar results in more educated participants (e.g., physicians) in different countries. Finally, future research could investigate whether the general findings hold across different types of visual aids (e.g., icon arrays, bar charts, and line plots), when visual aids are provided instead of rather than in addition to numerical information, and when visual aids differ in iconicity (i.e., when they are more or less abstract). In accord with the growing body of research, we predict that simple, well-designed visual aids will show substantial benefits in many situations, especially when communicating with less numerate individuals.

Author Contributions

All authors listed on the manuscript have contributed sufficiently to the project to be included as authors. All authors conceptualized the study, obtained funding, and wrote the paper. All authors approved the final version of the manuscript.

Acknowledgments

The current research was funded by the Ministerio de Economía y Competitividad (Spain) (PSI2011-22954 and PSI2014-51842-R), the National Science Foundation (USA) (SES-1253263), and the Swiss National Science Foundation (100014_140503). The authors declare independence from these funding agencies and do not have conflicts of interest including financial interests, activities, relationships, and affiliations.

References

Ancker, J. S., and Kaufman, D. (2007). Rethinking health numeracy: a multidisciplinary literature review. *J. Am. Med. Inform. Assoc.* 14, 713–721. doi: 10.1197/jamia.M2464

Ancker, J. S., Senathirajah, Y., Kukafka, R., and Starren, J. B. (2006). Design features of graphs in health risk communication: a systematic review. *J. Am. Med. Inform. Assoc.* 13, 608–618. doi: 10.1197/jamia.M2115

Ancker, J. S., Weber, E. U., and Kukafka, R. (2011). Effect of arrangement of stick figures on estimates of proportion in risk graphics. *Med. Decis. Mak.* 31, 143–150. doi: 10.1177/0272989X10369006

Apter, A. J., Paasche-Orlow, M. K., Remillard, J. T., Bennett, I. M., Ben-Joseph, E. P., Batista, R. M., et al. (2008). Numeracy and communication with patients: they are counting on us. *J. Gen. Intern. Med.* 23, 2117–2224. doi: 10.1007/s11606-008-0803-x

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol.* 23, 369–381. doi: 10.1002/acp.1460

Brase, G. L. (2014). The power of representation and interpretation: doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *J. Cogn. Psychol.* 26, 81–97. doi: 10.1080/20445911.2013.861840

Brase, G. L., Cosmides, L., and Tooby, J. (1998). Individuation, counting, and statistical inference: the role of frequency and whole-object representations in judgment under uncertainty. *J. Exp. Psychol. Gen.* 127, 3–21. doi: 10.1037/0096-3445.127.1.3

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., and Garcia-Retamero, R. (2012). Measuring risk literacy: the Berlin Numeracy Test. *Judgm. Decis. Mak.* 7, 25–47.

Cokely, E. T., Ghazal, S., and Garcia-Retamero, R. (2014). "Measuring numeracy," in *Numerical Reasoning in Judgments and Decision Making about Health*, eds B. L. Anderson and J. Schulkin (Cambridge: Cambridge University Press), 11–38.

Cokely, E. T., and Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: a protocol analysis and process model evaluation. *Judgm. Decis. Mak.* 4, 20–33.

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

Cox, D. S., Cox, A. D., Sturm, L., and Zimet, G. (2010). Behavioral interventions to increase HPV vaccination acceptability among mothers of young girls. *Health Psychol.* 29, 29–39. doi: 10.1037/a0016942

Davids, S. L., Schapira, M. M., McAuliffe, T. L., and Nattinger, A. B. (2004). Predictors of pessimistic breast cancer risk perceptions in a primary care population. *J. Gen. Intern. Med.* 19, 310–315. doi: 10.1111/j.1525-1497.2004.20801.x

Dunning, D., Heath, C., and Suls, J. M. (2004). Flawed self-assessment implications for health, education, and the workplace. *Psychol. Sci. Public Interest* 5, 69–106. doi: 10.1111/j.1529-1006.2004.00018.x

Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine: problems and opportunities," in *Judgment Under Uncertainty: Heuristics and Biases*, ed. D. Kahneman (Cambridge: Cambridge University Press), 249–267. doi: 10.1017/CBO9780511809477.019

Ehrlinger, J., and Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *J. Pers. Soc. Psychol.* 84, 5–17. doi: 10.1037/0022-3514.84.1.5

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., and Kruger, J. (2008). Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent. *Organ. Behav. Hum. Decis. Process.* 105, 98–121. doi: 10.1016/j.obhdp.2007.05.002

Ellis, K. M., Cokely, E. T., Ghazal, S., and Garcia-Retamero, R. (2014). Do people understand their home HIV test results? risk literacy and information search. *Proc. Hum. Fact. Ergon. Soc. Ann. Meet.* 58, 1323–1327. doi: 10.1177/1541931214581276

Fagerlin, A., Ubel, P. A., Smith, D. M., and Zikmund-Fisher, B. J. (2007). Making numbers matter: present and future research in risk communication. *Am. J. Health Behav.* 31, S47–S56. doi: 10.5993/ajhb.31.s1.7

Fagerlin, A., Wang, C., and Ubel, P. A. (2005). Reducing the influence of anecdotal reasoning on people's health care decisions: is a picture worth a thousand statistics? *Med. Decis. Mak.* 25, 398–405. doi: 10.1177/0272989X05278931

Feldman-Stewart, D., Brundage, M. D., and Zotov, V. (2007). Further insight into the perception of quantitative information: judgments of gist in treatment decisions. *Med. Decis. Mak.* 27, 34–43. doi: 10.1177/0272989X06297101

Feldman-Stewart, D., Kocovski, N., McConnell, B. A., Brundage, M. D., and Mackillop, W. J. (2000). Perception of quantitative information for treatment decisions. *Med. Decis. Mak.* 20, 228–238. doi: 10.1177/0272989X0002000208

Fischhoff, B., Brewer, N. T., and Downs, J. S. (2012). *Communicating Risks and Benefits: An Evidence Based User's Guide.* Silver Spring, MD: US Department of Health and Human Service, Food and Drug Administration.

Gaissmaier, W., Wegwarth, O., Skopec, D., Müller, A., Broschinski, S., and Politi, M. C. (2012). Numbers can be worth a thousand pictures: individual differences in understanding graphical and numerical representations of health-related information. *Health Psychol.* 31, 286–296. doi: 10.1037/a0024850

Galesic, M., and Garcia-Retamero, R. (2010). Statistical numeracy for health: a cross-cultural comparison with probabilistic national samples. *Arch. Intern. Med.* 170, 462–468. doi: 10.1001/archinternmed.2009.481

Galesic, M., and Garcia-Retamero, R. (2011a). Communicating consequences of risky behaviors: life expectancy versus risk of disease. *Patient Educ. Couns.* 82, 30–35. doi: 10.1016/j.pec.2010.02.008

Galesic, M., and Garcia-Retamero, R. (2011b). Do low-numeracy people avoid shared decision making? *Health Psychol.* 30, 336–341. doi: 10.1037/a0022723

Garcia-Retamero, R., Andrade, A., Sharit, J., and Ruiz, J. G. (2015). Is patient's numeracy related to physical and mental health? *Med. Decis. Mak.* 35, 501–511. doi: 10.1177/0272989X15578126

Garcia-Retamero, R., and Cokely, E. T. (2011). Effective communication of risks to young adults: using message framing and visual aids to increase condom use and STD screening. *J. Exp. Psychol. Appl.* 17, 270–287. doi: 10.1037/a0023677

Garcia-Retamero, R., and Cokely, E. T. (2013). Communicating health risks with visual aids. *Curr. Dir. Psychol. Sci.* 22, 392–399. doi: 10.1177/0963721413491570

Garcia-Retamero, R., and Cokely, E. T. (2014a). The influence of skills, message frame, and visual aids on prevention of sexually transmitted diseases. *J. Behav. Decis. Mak.* 27, 179–189. doi: 10.1002/bdm.1797

Garcia-Retamero, R., and Cokely, E. T. (2014b). "Using visual aids to help people with low numeracy make better decisions," in *Numerical Reasoning in Judgments and Decision Making about Health*, eds B. L. Anderson and J. Schulkin (Cambridge: Cambridge University Press), 153–174.

Garcia-Retamero, R., and Cokely, E. T. (2015). Simple but powerful health messages for increasing condom use in young adults. *J. Sex Res.* 52, 30–42. doi: 10.1080/00224499.2013.806647

Garcia-Retamero, R., and Galesic, M. (2009). Communicating treatment risk reduction to people with low numeracy skills: a cross-cultural comparison. *Am. J. Public Health* 99, 2196–2202. doi: 10.2105/AJPH.2009.160234

Garcia-Retamero, R., and Galesic, M. (2010a). How to reduce the effect of framing on messages about health. *J. Gen. Intern. Med.* 25, 1323–1329. doi: 10.1007/s11606-010-1484-9

Garcia-Retamero, R., and Galesic, M. (2010b). Who profits from visual aids: overcoming challenges in people's understanding of risks. *Soc. Sci. Med.* 70, 1019–1025. doi: 10.1016/j.socscimed.2009.11.031

Garcia-Retamero, R., and Galesic, M. (2011). Using plausible group sizes to communicate information about medical risks. *Patient Educ. Couns.* 84, 245–250. doi: 10.1016/j.pec.2010.07.027

Garcia-Retamero, R., and Galesic, M. (2013). *Transparent Communication of Health Risks: Overcoming Cultural Differences*. New York: Springer. doi: 10.1007/978-1-4614-4358-2

Garcia-Retamero, R., Galesic, M., and Gigerenzer, G. (2010). Do icon arrays help reduce denominator neglect? *Med. Decis. Mak.* 30, 672–684. doi: 10.1177/0272989X10369000

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

Garcia-Retamero, R., Wicki, B., Cokely, E. T., and Hanson, B. (2014). Factors predicting surgeons' preferred and actual roles in interactions with their patients. *Health Psychol.* 33, 920–928. doi: 10.1037/hea0000061

Gardner, P. H., McMillan, B., Raynor, D. K., Woolf, E., and Knapp, P. (2011). The effect of numeracy on the comprehension of information about medicines in users of a patient information website. *Patient Educ. Couns.* 83, 398–403. doi: 10.1016/j.pec.2011.05.006

Ghazal, S., Cokely, E. T., and Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: numeracy and metacognition. *Judgm. Decis. Mak.* 9, 15–34.

Gigerenzer, G. (2013). How I got started teaching physicians and judges risk literacy. *Appl. Cogn. Psych.* 28, 612–614. doi: 10.1002/acp.2980

Gigerenzer, G., Fiedler, K., and Olsson, H. (2012). "Rethinking cognitive biases and environmental consequences," in *Ecological Rationality: Intelligence in the World*, eds P. M. Todd, G. Gigerenzer, and the ABC Research Group (New York: Oxford University Press), 80–110.

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., and Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev.* 106, 425–430. doi: 10.1037/0033-295X.106.2.425

Gillan, D. J., Wickens, C. D., Hollands, J. G., and Carswell, C. M. (1998). Guidelines for presenting quantitative data in HFES publications. *Hum. Fact.* 40, 28–41.

Goodyear-Smith, F., Arroll, B., Chan, L., Jackson, R., Wells, S., and Kenealy, T. (2008). Patients prefer pictures to numbers to express cardiovascular benefit from treatment. *Ann. Fam. Med.* 6, 213–217. doi: 10.1370/afm.795

Griffin, D., and Brenner, L. (2004). "Perspectives on probability judgment calibration," in *Blackwell Handbook of Judgment and Decision Making*, eds D. J. Koehler and N. Harvey (Oxford: Blackwell), 177–199. doi: 10.1002/9780470752937.ch9

Gurmankin, A. D., Baron, J., and Armstrong, K. (2004). Intended message versus message received in hypothetical physician risk communications: exploring the gap. *Risk Anal.* 24, 1337–1347. doi: 10.1111/j.0272-4332.2004.00530.x

Hanoch, Y., Miron-Shatz, T., Cole, H., Himmelstein, M., and Federman, A. D. (2010). Choice, numeracy, and physicians-in-training performance: the case of medicare part D. *Health Psychol.* 29, 454–459. doi: 10.1037/a0019881

Hibbard, J. H., Peters, E., Dixon, A., and Tusler, M. (2007). Consumer competencies and the use of comparative quality information: it isn't just about literacy. *Med. Care Res. Rev.* 64,379–394. doi: 10.1177/1077558707301630

Hoffrage, U., and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Acad. Med.* 73, 538–540. doi: 10.1097/00001888-199805000-00024

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition* 84, 343–352. doi: 10.1016/S0010-0277(02)00050-1

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science* 290, 2261–2262. doi: 10.1126/science.290.5500.2261

Johnson, E. D., and Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Front. Psychol.* 6:938.

Kaplan, R. M., and Frosch, D. L. (2005). Decision making in medicine and health care. *Annu. Rev. Clin. Psychol.* 1, 525–556. doi: 10.1146/annurev.clinpsy.1.102803.144118

Kurzenhäuser, S., and Hoffrage, U. (2002). Teaching Bayesian reasoning: an evaluation of a classroom tutorial for medical students. *Med. Teach.* 24, 516–521. doi: 10.1080/0142159021000012540

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations. *Med. Decis. Mak.* 27, 696–713. doi: 10.1177/0272989X07307271

Lipkus, I. M., and Hollands, J. G. (1999). The visual communication of risk. *J. Natl. Cancer Inst. Monogr.* 25, 149–163. doi: 10.1093/oxfordjournals.jncimonographs.a024191

Lipkus, I. M., Samsa, G., and Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Med. Decis. Mak.* 21, 37–44. doi: 10.1177/0272989X0102100105

Mandel, D. R. (2014). Visual representation of rational belief revision: another look at the sleeping beauty problem. *Front. Psychol.* 5:1232. doi: 10.3389/fpsyg.2014.01232

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Front. Psychol.* 6:387. doi: 10.3389/fpsyg.2015.00387

Metcalfe, J., and Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychon. Bull. Rev.* 15, 174–179. doi: 10.3758/PBR.15.1.174

Nelson, T. O. (1990). Metamemory: a theoretical framework and new findings. *Psychol. Learn. Motiv.* 26, 125–173. doi: 10.1016/S0079-7421(08)60053-5

Okan, Y., Garcia-Retamero, R., Cokely, E. T., and Maldonado, A. (2012). Individual differences in graph literacy: overcoming denominator neglect in risk comprehension. *J. Behav. Decis. Mak.* 25, 390–401. doi: 10.1002/bdm.751

Okan, Y., Garcia-Retamero, R., Cokely, E. T., and Maldonado, A. (2015). Improving risk understanding across ability levels: encouraging active processing with dynamic icon arrays. *J. Exp. Psychol. Appl.* 21, 178–194. doi: 10.1037/xap0000045

Paling, J. (2003). Strategies to help patients understand risks. *BMJ* 327, 745–748. doi: 10.1136/bmj.327.7417.745

Peters, E. (2012). Beyond comprehension the role of numeracy in judgments and decisions. *Curr. Dir. Psychol. Sci.* 21, 31–35. doi: 10.1177/0963721411429960

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H., and Mertz, C. K. (2007). Less is more in presenting quality information to consumers. *Med. Care Res. Rev.* 64, 169–190. doi: 10.1177/1077558707301630

Peters, E., and Levin, I. P. (2008). Dissecting the risky-choice framing effect: numeracy as an individual-difference factor in weighting risky and riskless options. *Judgm. Decis. Mak.* 3, 435–448.

Peters, E., Vastfjall, D., Slovic, P., Mertz, C. K., Mazzocco, K., and Dickert, S. (2006). Numeracy and decision making. *Psychol. Sci.* 17, 407–413. doi: 10.1111/j.1467-9280.2006.01720.x

Petrova, D. G., Pligt, J., and Garcia-Retamero, R. (2014). Feeling the numbers: on the interplay between risk, affect, and numeracy. *J. Behav. Decis. Mak.* 27, 191–199. doi: 10.1002/bdm.1803

Portnoy, D. B., Roter, D., and Erby, L. H. (2010). The role of numeracy on client knowledge in BRCA genetic counseling. *Patient Educ. Couns.* 81, 131–136. doi: 10.1016/j.pec.2009.09.036

Reyna, V. F. (2004). How people make decisions that involve risk: a dual-processes approach. *Curr. Dir. Psychol. Sci.* 13, 60–66. doi: 10.1111/j.0963-7214.2004.00275.x

Reyna, V. F., and Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learn. Individ. Differ.* 18, 89–107. doi: 10.1016/j.lindif.2007.03.011

Reyna, V. F., Nelson, W. L., Han, P. K., and Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychol. Bull.* 135, 943–973. doi: 10.1037/a0017327

Rothman, R. L., Housam, R., Weiss, H., Davis, D., Gregory, R., Gebretsadik, T., et al. (2006). Patient understanding of food labels: the role of literacy and numeracy. *Am. J. Prev. Med.* 31, 391–398. doi: 10.1016/j.amepre.2006.07.025

Schirillo, J. A., and Stone, E. R. (2005). The greater ability of graphical versus numerical displays to increase risk avoidance involves a common mechanism. *Risk Anal.* 25, 555–566. doi: 10.1111/j.1539-6924.2005.00624.x

Schraw, G. (2010). Measuring self-regulation in computer-based learning environments. *Educ. Psychol.* 45, 258–266. doi: 10.1080/00461520.2010.515936

Schwartz, L. M., Woloshin, S., Black, W. C., and Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Ann. Intern. Med.* 127, 966–972. doi: 10.7326/0003-4819-127-11-199712010-00003

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science* 333, 1393–1400. doi: 10.1126/science.1191181

Stankov, L. (2000). Structural extensions of a hierarchical view on human cognitive abilities. *Learn. Individ. Differ.* 12, 35–51. doi: 10.1016/S1041-6080(00)00037-6

Stankov, L., and Lee, J. (2008). Confidence and cognitive test performance. *J. Educ. Psychol.* 100, 961–976. doi: 10.1037/a0012546

Stone, E. R., Sieck, W. R., Bull, B. E., Yates, F. J., Parks, S. C., and Rush, C. J. (2003). Foreground: background salience: explaining the effects of graphical displays on risk avoidance. *Organ. Behav. Hum. Decis. Process.* 90, 19–36. doi: 10.1016/S0749-5978(03)00003-7

Thompson, V. A., Prowse Turner, J. A., and Pennycook, G. (2011). Intuition, reason, and metacognition. *Cogn. Psychol.* 63, 107–140. doi: 10.1016/j.cogpsych.2011.06.001

Trevena, L., Zikmund-Fisher, B., Edwards, A., Gaissmaier, W., Galesic, M., Han, P., et al. (2012). "Presenting probabilities," in *Update of the International Patient Decision Aids Standards (IPDAS) Collaboration's Background Document*, eds R. Volk and H. Llewellyn-Thomas. Available at: http://ipdas.ohri.ca/IPDAS-Chapter-C.pdf

Waters, E. A., Weinstein, N. D., Colditz, G. A., and Emmons, K. M. (2007). Reducing aversion to side effects in preventive medical treatment decisions. *J. Exp. Psychol. Appl.* 13, 11–21. doi: 10.1037/1076-898X.13.1.11

Weinfurt, K. P., Castel, L. D., Li, Y., Sulmasy, D. P., Balshem, A. M., Benson, A. B., et al. (2003). The correlation between patient characteristics and expectations of benefit from phase I clinical trials. *Cancer* 98, 166–175. doi: 10.1002/cncr.11483

Zikmund-Fisher, B. J. (2015). Stories of MDM: from a conversation to a career of making less data more useful. *Med. Decis. Mak.* 35, 1–3. doi: 10.1177/0272989X14563576

Zikmund-Fisher, B. J., Fagerlin, A., and Ubel, P. A. (2008a). Improving understanding of adjuvant therapy options by using simpler risk graphics. *Cancer* 113, 3382–3390. doi: 10.1002/cncr.23959

Zikmund-Fisher, B. J., Ubel, P. A., Smith, D. M., Derry, H. A., McClure, J. B., Stark, A., et al. (2008b). Communicating side effect risks in a tamoxifen prophylaxis decision aid: the debiasing influence of pictographs. *Patient Educ. Couns.* 73, 209–214. doi: 10.1016/j.pec.2008.05.010

Zikmund-Fisher, B. J., Fagerlin, A., and Ubel, P. A. (2010). A demonstration of "less can be more" in risk graphics. *Med. Decis. Mak.* 30, 661–671. doi: 10.1177/0272989X10364244

Zikmund-Fisher, B. J., Witteman, H. O., Dickson, M., Fuhrel-Forbis, A., Kahn, V. C., Exe, N. L., et al. (2014). Blocks, ovals, or people? icon type affects risk perceptions and recall of pictographs. *Med. Decis. Mak.* 34, 443–453. doi: 10.1177/0272989X13511706

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 August 2015
doi: 10.3389/fpsyg.2015.00939

Toward an ecological analysis of Bayesian inferences: how task characteristics influence responses

Sebastian Hafenbrädl and Ulrich Hoffrage*

Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland

In research on Bayesian inferences, the specific tasks, with their narratives and characteristics, are typically seen as exchangeable vehicles that merely transport the structure of the problem to research participants. In the present paper, we explore whether, and possibly how, task characteristics that are usually ignored influence participants' responses in these tasks. We focus on both quantitative dimensions of the tasks, such as their base rates, hit rates, and false-alarm rates, as well as qualitative characteristics, such as whether the task involves a norm violation or not, whether the stakes are high or low, and whether the focus is on the individual case or on the numbers. Using a data set of 19 different tasks presented to 500 different participants who provided a total of 1,773 responses, we analyze these responses in two ways: first, on the level of the numerical estimates themselves, and second, on the level of various response strategies, Bayesian and non-Bayesian, that might have produced the estimates. We identified various contingencies, and most of the task characteristics had an influence on participants' responses. Typically, this influence has been stronger when the numerical information in the tasks was presented in terms of probabilities or percentages, compared to natural frequencies – and this effect cannot be fully explained by a higher proportion of Bayesian responses when natural frequencies were used. One characteristic that did not seem to influence participants' response strategy was the numerical value of the Bayesian solution itself. Our exploratory study is a first step toward an ecological analysis of Bayesian inferences, and highlights new avenues for future research.

Keywords: Bayesian inference, updating beliefs, ecological analysis, task characteristics, base rate, signal-detection, representation format, natural frequencies

OPEN ACCESS

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
Gary L. Brase, Kansas State University, USA
Elisabet Tubau, Universitat de Barcelona, Spain

****Correspondence:***
Sebastian Hafenbrädl, Faculty of Business and Economics, University of Lausanne, Batiment Internef, Dorigny, CH-1015 Lausanne, Switzerland
sebastian.hafenbraedl@unil.ch

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *18 May 2015*
Accepted: *22 June 2015*
Published: *04 August 2015*

Citation:
Hafenbrädl S and Hoffrage U (2015) Toward an ecological analysis of Bayesian inferences: how task characteristics influence responses. Front. Psychol. 6:939. doi: 10.3389/fpsyg.2015.00939

Introduction

A woman receives a positive HIV test—what is the probability that she is infected? An eyewitness claims that she saw a blue cab involved in an accident—what is the probability that the cab was actually blue? A potential customer asks for a second sales presentation—what is the probability that he will ultimately place an order? Even though these questions come from different domains, they all share the same underlying structure: an individual receives new diagnostic information and wants to update her beliefs accordingly. Tasks that provide (a) information about prior probabilities of some hypotheses, (b) information that new evidence is available, and (c) information about the probabilities of such new evidence under various conditions, are called Bayesian inference

problems and their solution can be calculated using Bayes' rule. For more than 50 years, researchers have been interested in the psychological processes individuals deploy to solve such problems as well as how to help individuals to solve such problems more effectively (Mandel, 2014, 2015; Johnson and Tubau, 2015; McNair, 2015).

Sirota et al. (2015) pointed out that Bayesian reasoning is not restricted to textbook problems, and that there is a wide range of situations that call for Bayesian reasoning "in the wild," that is, in real life contexts in which information is usually *not* provided in numerical form and in which people (and animals) nevertheless have to behave after some events occurred or after some information became known (see also Griffiths and Tenenbaum, 2006). This distinction is akin of Hertwig et al.'s (2004) distinction between *decisions-from-description* and *decisions-from-experience*. In a similar vein, Mandel (2014) called for adopting a wider perspective and for studying Bayesian reasoning in domains other than textbook problems. In the present paper, we do not follow this call, and we analyze, as most researchers on Bayesian reasoning do, people's responses to textbook problems. Yet, we aim at going beyond the usual treatment of such problems. Usually, the content of a given task is just regarded as decoration—what is important is that the task has a certain structure and that this structure and the information given in the task qualify it as a Bayesian inference task for which Bayes' rule, as a "content-blind" norm, provides the solution (Gigerenzer, 1996). We question what often seems to be taken for granted, namely that task content does not matter and is exchangeable. This avenue does not lead us into the wild, but it leads us into white territory from the viewpoint of classic textbook problem analysis. We seek to explore the effect of task dimensions that are usually ignored.

We are not the first to challenge the tacit assumption that task content is decorative and can be ignored as long as it serves its purpose, namely to convey what the structure is and which normative principle applies. For instance, Cosmides (1989) argued that the content and context of the task used to study deductive reasoning matters: while a Wason selection task with an abstract content yields very few normatively correct responses, people's ability to correctly apply the modus tolens increased dramatically when the rule that needed to be checked was formulated as a social contract—even though this was irrelevant from a logical point of view. Another example is Krynski and Tenenbaum's (2007) finding that performance in a Bayesian reasoning task depends on verbal content, specifically, whether a reason for a false-alarm in a medical test has been given ("the presence of a benign cyste") or not. Note that providing participants with an alternative cause for a positive test is irrelevant from a normative point of view because it does not affect the false-alarm rate. In other words, the false-alarm rates in both versions, with and without reason for the false-alarm, were the same. However, providing a reason boosted the proportion of Bayesian answers from about 25% to about 45%—which is, according to Johnson and Tubau (2015), "some of the highest performance reported with normalized data in the absence of visual cues." To provide one more example, Mellers and McGraw (1999) hypothesized that the beneficial effect of natural frequencies (for an explanation of this concept, see below) is minimized for tasks with a high base rate, which amounts to saying that the usage of a Bayesian response strategy in the probability/percentage version and in the natural frequency version is differentially affected by the base rate stated in the problem. Note that the claim is not that the Bayesian solution depends on the base rate—this is trivial and follows from Bayes' rule. Rather the claim is empirical in nature, namely that a participant's chance of answering with the Bayesian solution does depend on the base rate. Mellers and McGraw (1999) provided supportive evidence for their interaction hypothesis, and when we tested it with our own data, we could confirm that the pattern of results for the cab problem (which Mellers and McGraw used) seemed indeed to be special, but we could not obtain supportive evidence for the hypothesized interaction in general (Gigerenzer and Hoffrage, 1999).

Our research question is directly in line with these three examples: are there characteristics of Bayesian textbook problems—and if so, which—that influence participants' responses? Note that this investigation conceives participants' responses to Bayesian inference tasks as a function of task characteristics and can thus be considered as an example of how strategy usage depends on ecological dimensions (Todd et al., 2012).

Materials and Methods

Databasis

To explore how characteristics of Bayesian inference tasks influence responses and the usage of response strategies, we reanalyzed data that was obtained by prior research. In particular, we pooled the data from Hoffrage et al. (2015) and the data from Study 1 of Gigerenzer and Hoffrage (1995). Our pooled data set consists of 19 different tasks (4 tasks from Hoffrage et al., 2015 and 15 tasks from Gigerenzer and Hoffrage, 1995), presented to a total of 500 different participants who provided 1,773 responses. **Table 1** gives an overview of these tasks and how they score on various quantitative and qualitative dimensions (which will be introduced in more detail below).

Tasks 13, 15, 17, and 18 have been taken from Hoffrage et al. (2015; for the full descriptions, see their introduction, their Table 1, and their Appendix). The 440 participants who worked on these tasks were 259 undergraduate students of a business school and 181 managers in their role as students in an Executive MBA program. For each of the four tasks, two versions were constructed, one in which the information was presented in percentages and one in which natural frequencies were used. Each of the participants responded to two different tasks, either two percentage versions, or two natural frequency versions; in other words, representation format (henceforth the label for his variable) has been manipulated between-subjects.

Natural frequencies are the tallies in a natural sample in which hit rate and false-alarm rate are not normalized with respect to base rates (see Hoffrage et al., 2002 and Gigerenzer and Hoffrage, 2007; for an example of how probability information

TABLE 1 | The 19 tasks used in the present analysis, ordered according to base rates.

	Task Content		Quantitative Dimensions					Qualitative Dimensions			
Task	**Hypothesis (H)**	**Data (D)**	**Base rate (Br)**	**Hit rate (Hr)**	**False-alarm rate (F)**	**Size of Reference Class**	**Bayesian response (Bay)**	**Norm deviation (N)**	**Stakes (S)**	**Main focus (M)**	**Number of responses**
1	Pimp	Wearing a Rolex	0.005	80	0.05	1000000	7.41				60
2	HIV infection	HIV-test positive	0.01	100	0.1	1000000	9.09	1	1	1	60
3	Heroin addict	Fresh needle prick	0.01	100	0.1	100000	5	1	1	1	60
4	Committing suicide	Professor	0.024	15	12	1000000	0.03	1	1		60
5	Prenatal damage in child	German measles in mother	0.21	47.6	0.5	10000	16.7	1	1	1	60
6	Breast cancer	Mammography positive	1	80	9.6	1000	7.77	1	1	1	60
7	Car accident	Driver drunk	1	55	5.05	10000	9.91	1	1		60
8	Pregnant	Pregnancy test positive	2	95	0.51	1000	79.17	0.5	1	1	53
9	Accident on way to school	Child lives in urban area	3	90	40	1000	6.51	1	1		60
10	Bad posture in child	Heavy books carried daily	5	40	20	1000	9.52	1	0.5		60
11	Active feminist	Bank teller	5	0.4	2.11	100000	0.99				60
12	Blue cab	Eyewitness says “Blue”	15	80	20	100	41.38		0.5		60
13	Incorrect tax return	Error detected	20	30	10	1000	42.86	1	1	1	218
14	Choosing course in economics	Career-oriented	30	70	50	1000	37.5				60
15	Supplier A	Material defective	30	15	10	1000	39.13		1		221
16	Admission to school	Particular placement test result	36	75	20	1000	67.84				60
17	Get contract	Invited to second presentation	60	70	50	100	67.74		1	1	222
18	Produced in Ohio	Container Defective	60	5	10	1000	42.86		1		219
19	Red ball	Marked with star	80	75	25	500	92.31				60

Tasks 13, 15, 17, and 18 have been taken from Hoffrage et al. (2015), and the others from Gigerenzer and Hoffrage (1995). For each task, two versions were constructed, one in which the information was presented in probabilities (Gigerenzer and Hoffrage, 1995) or percentages (Hoffrage et al., 2015) and one in which natural frequencies were used. The natural frequency versions can be constructed by applying the base rate to the size of the reference class. For instance, applying the base rate of 1% in Task 6 to the reference class of size 1,000 yields 10 out of 1,000. Applying the hit rate of 80% to these 10 yields 8 out of 10, and applying the false-alarm rate of 9.6 to the remaining 990 yields 95 out of 990. The three qualitative dimensions Norm Deviation (N), Stakes (S), and Main Focus (M) are explained in the text. The numbers of responses are aggregated across both task versions.

can be translated into natural frequencies, see the caption of **Table 1**). Natural Frequencies have proven to facilitate diagnostic inferences in laypeople (Gigerenzer and Hoffrage, 1995), advanced medical students and advanced law students (Hoffrage et al., 2000), patients (Garcia-Retamero and Hoffrage, 2013), physicians (Hoffrage and Gigerenzer, 1998), and managers and management students (Hoffrage et al., 2015). For a discussion about when and why natural frequencies are effective, see Gigerenzer and Hoffrage (2007), Brase (2008), Hill and Brase (2012), Brase and Hill (2015), and Johnson and Tubau (2015).

The remaining 15 tasks were taken from Gigerenzer and Hoffrage (1995; see Table 2, p. 293). In this study, four versions were constructed per task, but for the present re-analysis, we will only use two versions, namely the probability version and the natural frequency version of what Gigerenzer and Hoffrage (1995) called the standard menu. The information provided in the standard menu is displayed in **Table 1** (the two other versions involving the so-called short menu, which provides the information about the conjunctions *D*&*H* and *D*&–*H*, either in probabilities or in natural frequencies, are not included in the present re-analysis). Each of the 60 participants of Gigerenzer and Hoffrage (1995) received all 15 tasks in two versions. For 30 participants, these were probabilities, standard menu and natural frequencies, short menu, and for the other 30 participants, these were probabilities, short menu and natural frequencies, standard menu. The experiment took place in two sessions, most of them one week apart from each other. For each participant, half of the tasks in the first session were presented in one version, and the other half were presented in the other version. During a given session, a given participant has seen each task only once, that is, the two versions of the same task were given in different sessions. For the present re-analysis, which ignores all responses in the short menu version, this implies that we used a between-subject design: 30 participants responded to 15 tasks, each with information presented in terms of probabilities (seven tasks in one session and eight tasks in another session, one week apart from each other), and 30 other participants did the same, except that they were presented with the natural frequency versions of the same 15 tasks.

While both studies used a natural frequency condition, the condition with normalized information differed between the studies: Hoffrage et al. (2015) used percentages for their four tasks, and Gigerenzer and Hoffrage (1995) used probabilities for their 15 tasks. According to Gigerenzer and Hoffrage's (1995) analysis (result 7, p. 689), this difference should not have an effect on Bayesian performance—a theoretical result that was confirmed by their own data and in numerous studies of other authors since then. We will thus pool the data from these two studies, and we will, henceforth, refer to this condition as the probability/percentage condition.

Task Characteristics: Quantitative Dimensions

We will now introduce some candidate predictor variables that may explain some variance, across tasks, of participants' responses. One obvious dimension along which the tasks vary is the numeric information: the base rate, the hit rate and the false-alarm rate. Note that the third example given in our introduction (Mellers and McGraw, 1999) was of this kind: the authors argued that the chance of responding with the Bayesian solution (which must not be confused with the Bayesian solution itself!) is affected by whether the base rate is high or low. There are some observations about this set of three quantitative variables that we can make already before looking at the participants' data. First, the prior probability (i.e., base rate) is linked to the posterior probability: in our set of 19 tasks, the correlation is 0.76. The fact that this correlation is positive and substantial is trivial as the following analysis shows. Consider the so-called odds version of Bayes' rule, which can be read as a division of two equations (more precisely: after the posterior odds ratio has been extended by $p(D)/p(D)$, the four numerators constitute one equation and the four denominators the other one):

$$\underbrace{\frac{p(H|D)}{p(-H|D)}}_{\textit{posterior odds ratio}} = \underbrace{\frac{p(H)}{p(-H)}}_{\textit{prior odds ratio}} \times \underbrace{\frac{p(D|H)}{p(D|-H)}}_{\textit{likelihood ratio}} \qquad (1)$$

Equation 1 has the following implications: (a) if the likelihood ratio equals 1—which means that the data D is not at all diagnostic—then the posterior odds ratio is identical to the prior odds ratio, (b) if the likelihood ratio is larger than 1—which is usually the case and which is also the case for 17 of our 19 tasks—then the posterior odds ratio exceeds the prior odds ratio, and (c) the posterior odds ratio is a linear function of the prior odds ratio, with the likelihood ratio as a constant. Hence, one should expect a positive correlation between prior probability and posterior probability (although this link could be offset in a sample of tasks by some correlation patterns between the likelihood ratios and the corresponding prior odds in these tasks). For the sake of completeness, we want to mention that the correlation between base rate and the Bayesian solution (recall, 0.76) was found to be substantially higher than any other correlations that included the hit rate (Hr) and the false-alarm rate (F): corr (Bay*Hr) = 0.16, corr (Bay*F) = 0.34, corr (Br*Hr) = −0.15, corr (Br*F) = 0.51 and corr (Hr*F) = 0.14.

Task Characteristics: Qualitative Dimensions

Besides these quantitative dimensions, we categorized the 19 tasks along three qualitative dimensions. Note that the second example in our introduction (Krynski and Tenenbaum, 2007) was of this kind: these authors demonstrated that the chances of responding with the Bayesian solution is affected by whether or not a reason for the existence of false-alarms is given. We agree that this is an interesting variable and we embrace Johnson and Tubau's (2015) problem solving approach to Bayesian reasoning that can account for why providing a reason facilitates Bayesian performance. We would have appreciated to also include this variable in the present analysis—yet, none of the 19 tasks provided such a reason, and accounting for variance on a criterion variable is pointless if there is no variance on the predictor variable. Fortunately, we were able to identify three other variables as meaningful and interesting for our purpose at hand.

The first variable is henceforth referred to as *norm deviation*. It denotes whether the focal hypothesis constitutes a deviation from

a norm, in the sense that H can be considered an exception or something unusual that requires specific attention, whereas $-H$ can be considered the normal case. To illustrate this variable with some examples from our set of 19 tasks, norm deviation has been coded as 1 for breast cancer (vs. non-breast cancer), HIV infection (vs. no infection), and incorrect tax report (vs. correct tax report). Note that such tasks can be conceived as signal-detection tasks: signals (or data, D) are used to detect norm deviations (or H). In contrast, norm deviation was coded as 0 for tasks in which it seemed to be hard, if not impossible, to say which was the normal case; for instance, red ball (vs. blue ball), blue cap (vs. green cap), or supplier A (vs. supplier B; for more examples, see **Table 1**).

Our second qualitative variable, henceforth referred to as *stakes*, denotes whether being in the state of H or $-H$ makes a big difference (e.g., being infected with HIV, having an accident, or causing a prenatal damage has been coded as 1) or whether the stakes are either relatively low or not specified in the task description (e.g., drawing a red ball from an urn, being an active feminist, or choosing a course in economics has been coded as 0; **Table 1**).

Finally, our third qualitative variable is the *main focus* of the task. The main focus can either be on the individual case or on the numbers involved. For many tasks, the context story makes it clear that the central question is whether some individual or protagonist is in the state of H or $-H$, and the numbers given in the task description mainly serve the purpose of determining whether, given the observed data, this individual case should be treated as if H (vs. $-H$) were true. Examples include the questions of whether a specific woman (with a positive mammogram) has breast cancer or not, or whether a specific man (with fresh needle pricks) is a heroin addict or not. For such tasks, this variable has been coded as 1. In contrast, it was coded as 0 for tasks in which the main focus was on the relationship between the data D and the hypothesis H, in particular, on the posterior probability (or the corresponding relative frequency). The individual case is rather in the background and serves as an illustration. Examples include the Varden Soap task in which the vice president for production is not at all interested in the treatment of an individual soap container that was identified as defective, but in which he adopts a long run perspective and wonders about a fair allocation of costs between the two production facilities Ohio (H) and Virginia ($-H$) (see the appendix of Hoffrage et al., 2015).

Dependent Variables

To find out how the quantitative and qualitative task characteristics can account for variance on participants' responses, we analyze these responses in two ways. Specifically, our first dependent variable is the participants' numerical estimate, which is continuous and comes in form of probabilities, percentages, or ratios, ranging from 0 to 100%. Our second dependent variable is the cognitive strategy a participant used to combine the given numbers (e.g., whether s/he provided the hit rate as a response). This is a categorical variable with as many levels as there are strategies, but it can also be seen as a vector of mutually exclusive binary (dummy) variables, each of which coded as present (i.e., with a value of 1) if a certain strategy is used, and which yields, aggregated across responses, proportions.

Two of these cognitive strategies that we used as a model in our analyses below are the base rate and the hit rate. The base rate is identical to the normative (i.e., the Bayesian) solution if the likelihood ratio is 1, that is, if the diagnostic information is not at all diagnostic—which is the case if the hit rate and the false-alarm rate are identical. Providing the hit rate as a response has been referred to as the "inverse fallacy" (Koehler, 1996; Villejoubert and Mandel, 2002), the "Fisherian algorithm" (Gigerenzer and Hoffrage, 1995) or the "conversion error" (Wolfe, 1995), and it has been accounted for by the representativeness heuristic (Kahneman and Tversky, 1972) or the "confusion hypothesis" (Macchi, 1995). Providing the false-alarm rate as a response happened in only 1.6% of the cases and so we decided to omit the results for this strategy in our analyses below (in Hoffrage et al., 2015, this occurred in 3.2% of the responses, see their Table 2; and it did not even pass the 1% threshold in Gigerenzer and Hoffrage, 1995, see their Table 3).

The two other cognitive strategies that we used are the Bayesian, as the normative response strategy, and the joint occurrence, which is the probability (or percentage) of cases in which both the data (D) are present and the hypothesis (H) is true: $p(D\&H)$. This number can easily be calculated by multiplying the base rate and the hit rate ($p(D\&H) = \mathrm{p(D|H)}^{*}\mathrm{p(H)}$; or by applying the hit rate information to the base rate of the focal hypotheses, e.g., 10 out of 1,000 women have breast cancer and 8 out 10 woman with breast cancer test positive, hence 8 out of 1,000 have breast cancer *and* test positive, see **Table 1**). Joint occurrence is the numerator of Bayes' rule, and given that $p(H|D) = p(D\&H)/p(D)$, it can be seen as a step toward the Bayesian solution that falls short of carrying the computation to the end (see Johnson and Tubau, 2015). While we only classified responses as stemming from the base-rate or the hit-rate strategy when the responses were the exact values of the base rate or the hit rate, respectively, we used a more lenient criterion for those strategies for which the number could not simply be read off but had to be computed. Specifically, we classified responses as Bayesian or as stemming from the joint occurrence strategy when the responses were in the range of $\pm 1\%$ point from the value of the Bayesian solution or joint occurrence, respectively.

Gigerenzer and Hoffrage (1995) identified a wide range of other strategies, some of them were very exotic, have rarely been used, and basically reveal that participants had no clue and combined the numbers in an arbitrary and/or unreliable way. Such attempts come close to guessing, and many participants said right away that they simply guessed, without having been able to say in which way they used the numbers exactly. Whether 'guessing' deserves being labeled as a strategy is a matter of taste—pragmatically, that is, from a modeling point of view, it is useless as 'guessing' does not allow one to make predictions and to calculate goodness-of-fit measures. In sum, we restricted the report of our results to four cognitive strategies—Bayes, base rate, hit rate, joint occurrence—each of which made a precise point prediction. Based on the previous literature (in particular Gigerenzer and Hoffrage, 1995, and Hoffrage et al., 2015), these were the most frequently used strategies.

Results

We structure the report of our results as follows: first, we analyze how the quantitative variables defined above are related to the qualitative dimensions of the tasks, and, second, how the qualitative variables are related to each other. Note that these analyses are conducted without any participant data. Third, we will present an overview of our data that comes close to presenting the raw data, thereby comprising all essential variables of the present analysis. Fourth, we will report how the quantitative and, fifth, how the qualitative task dimensions affect the numerical estimates. These analyses ignore process data and do not take into account whether a participant used a particular cognitive strategy, for instance, gave the hit rate as a response. Subsequently, we will turn to those 52.3% of the responses that have been identified as stemming from one of the most prominent strategies. For this subset of our data, we will analyze how, sixth, the representation format, seventh, the quantitative, and, eigth, the qualitative dimensions affect the usage of cognitive strategies.

How are the Quantitative Dimensions Related to the Qualitative Dimensions?

In many real world contexts it may be the base rates that determine which category stands out and attracts special attention. In fact, for our 19 tasks, the correlation between base rate and norm deviation is −0.64 [average base rate for tasks with norm deviation coded as 1 and 0 is 3.4% and 35.1%, respectively, $t(16) = 3.43$, $p = 0.004$]. Similarly, the correlation between the Bayesian solution and norm deviation is −0.56 [average Bayesian solution for tasks with norm deviation coded as 1 and 0 is 11.9 and 44.1%, respectively, $t(16) = 3.06$, $p = 0.008$], that is, tasks for which *H* constitutes a norm deviation tend to have lower Bayesian solutions. Moreover, it can be expected that for these tasks (for which *H* can be seen as a norm deviation), the stakes are high. If we consider natural catastrophes, diseases, crimes, fraud, or failure of technical systems, then we find that these are not only rare events and norm deviations, but that there are usually also high incentives to detect them early in order to be able to intervene and to prevent the worst. In other words, stakes are high. Hence not surprisingly, Pleskac and Hertwig (2014) observed that for many events in the real world, probabilities and utilities are negatively correlated: the lower the probability of events, the higher their magnitude in utility terms, either as a cost (e.g., earthquakes with higher severities are less likely), or as a benefit (e.g., higher stakes lotteries are less likely to be won). Consistent with Pleskac and Hertwig (2014), the correlation between base rate and stakes that we observed in our set of 19 tasks is negative [−0.26; the average base rates for high stake tasks is 14.8% and for low stake tasks it is 30.2%; $t(15) = 1.13$, $p = 0.27$]. In turn, the Bayesian solutions for high stakes tasks are also lower (27.2%), than for low stakes tasks [41.2%, $t(15) = 0.86$, $p = 0.40$]. Also our third qualitative variable is correlated with some of the quantitative variables: even though the base rate for problems in which the main focus is on the individual is lower than when the main focus is on the numbers [11.9 vs. 22.1, $t(17) = 0.86$, $p = 0.40$], this does not translate into differences in the Bayesian solution [32.6 vs. 29.6, $t(17) = -0.21$, $p = 0.83$]. This pattern can be explained by a combination of both smaller false-alarm rates [10.1 vs. 17.9, $t(17) = 1.0$, $p = 0.33$] and higher hit rates [74.7 vs. 50.0, $t(17) = -1.66$, $p = 0.12$].

How are the Qualitative Dimensions Related to Each Other?

All correlations in the triangle of qualitative variables are substantial and significant. The one between norm deviation and stakes is 0.62 ($p = 0.005$), that is, in tasks centering on norm deviations and abnormal cases, stakes tend to be high. The correlation between norm deviation and main focus is 0.45 ($p = 0.05$), that is, tasks about norm deviations tend to focus on the individual case. Moreover, the correlation between stakes and main focus is 0.55 ($p = 0.01$), that is, problems involving high stakes tend to focus on the individual case.

The results reported so far did not contain any participant responses and could hence have been reported before the first participant has shown up. Nevertheless, these are empirical findings that capture aspects of the statistical structure of Bayesian tasks. We will now turn to participants' responses.

How are Participants' Responses Distributed in the 19 Tasks?

Figure 1 displays the 19 tasks listed in **Table 1**. It thereby uses the same order, namely the one established by the base rates, and the identification numbers in **Table 1** correspond to those in **Figure 1**. This figure comes close to a presentation of the raw data. It visualizes, for each task, all variables that are included in the present analyses: the two sets of predictor variables (quantitative and qualitative task dimensions), and the two kinds of dependent variables (numerical estimates and response strategies). The quantitative dimensions of the task are included as lines that represent the numerical values of the base rate (Br), of the hit rate (Hr), of the false-alarm rate (F), and also of the Bayesian solution (Bay). The letters that stand for the three qualitative dimensions introduced above—norm deviation (N), stakes (S), and main focus (M)—indicate that the corresponding variable has been coded as "1" (absence of a letter for a given task indicates that the dimension has been coded as "0"). On the side of the dependent variables, the figure displays the distribution of numerical estimates, highlighting the estimates that correspond to specific response strategies in vertical bars, while all other responses that could not be assigned to one of the strategies that we selected for this analysis are visualized in a horizontal bar. The height of the vertical bars depicts the relative frequency of response strategy usage.

The advantage of this kind of data representation is, at the same time, its disadvantage. The figure contains a lot of information and is very detailed. In the subsequent sections we will hence focus on specific effects that the predictor variables exert on the dependent variables, that is, we split the data into subgroups and aggregate them so that some effects become better visible.

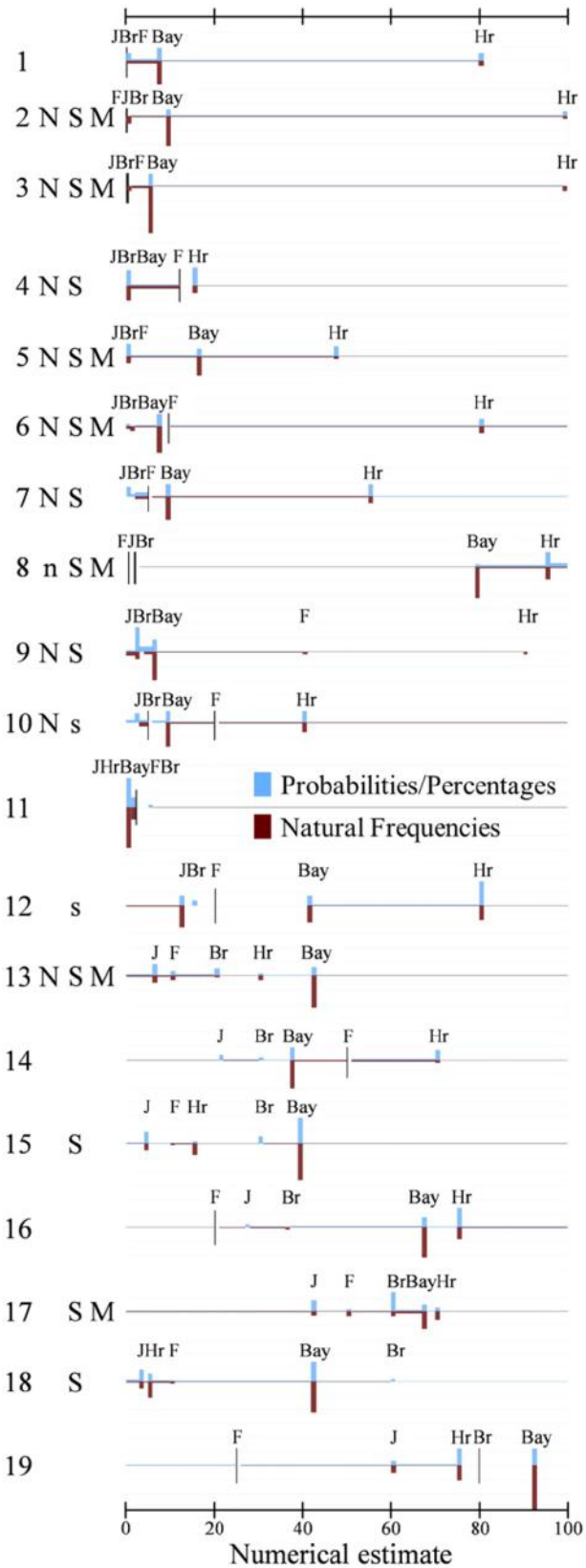

FIGURE 1 | Continued

FIGURE 1 | Continued

Distribution of numerical estimates for the 19 tasks compiled in Table 1 split by representation format. The 19 tasks are displayed below each other, and sorted by their base rate. Responses to task versions in which numbers were presented in terms of probabilities or percentages are plotted, in blue color, above the *x*-axes, and responses to natural frequency versions are plotted, in red color, below the *x*-axes. The figure does not highlight all numerical estimates, but only those for the following five strategies: Bayes (Bay), base rate (Br), hit rate (Hr), false-alarm rate (F) and joint-occurrence (J). The *y*-axes of Figure plots the frequency of use of these strategies, that is, of giving the corresponding number as an estimate (one gridline indicates 20% points) against the number that correspond to these strategies (on the *x*-axes). These five strategies together account for a total of 53.9% of the responses; the remaining 46.1% are distributed across the six intervals that are defined by the five numerical estimates of the five strategies. The widths of the bars for the strategies was set to be 1% and the heights of the intervals between the strategies (or between the strategies and the end points of the scales) were chosen such that equal areas amount to equal percentages of participants responding with the corresponding numerical estimate (be it the precise number of a given strategy or any estimate that falls within a given interval). For some tasks, the numerical estimates corresponding to some strategies yielded a value between 0 and 1, so that our chosen resolution would require to plot them behind each other, with overlapping bars (and with the consequence that the total areas would no longer be constant across tasks). For these tasks, we stacked the areas corresponding to the involved strategies on top of each other, so that the bars gained in height and were comprised of different strategy users. Specifically, for Task 1 (probability version, *p*) the bar at 0 represents: 1 Br; Task 1 (natural frequency version, nf): 3J; for Task 4p: 4J and 2Bay; Task 4nf: 1J, 1Br, and 4Bay; Task 5nf: 1J and 2Br; Task 11p: 1J, 6Hr, and 5Bay; Task 11nf: 1J, 3Hr, and 13Bay. Moreover, for 10 of the tasks from Gigerenzer and Hoffrage (1999), a total of 46 responses (corresponding to 9.6% of all responses for these tasks) could not be displayed because they fell between two adjacent responses of different strategies that were too close to each other to allow for graphical representation. In addition, the figure contains the classification for the three qualitative variables introduced in section "Materials and Methods": Norm deviation (N), Stakes (S), and Main focus (M). Capital letters denote that the hypothesized event constitutes a norm deviation, that stakes were high and that the main focus was on the individual case; absent letters denote the opposite; and letters in lower case denote that we could not agree how to code this variable (for instance, is being pregnant a norm deviation or not?). When computing the correlations or the regressions that are reported in the text, such unclear cases have been coded with 0.5, and when reporting the relative frequencies of strategy usage, the results for this variable level have been omitted.

How do the Quantitative Dimensions Affect the Numerical Estimates?

To see how the numbers given in the task affect the numerical estimates of the participants, we choose a data representation that combines (a) scatter-plots in which each dot denotes the average numerical estimate for a given task and information representation format, with (b) marginal effects from regression analysis and their corresponding confidence intervals (**Figure 2**).

Figure 2A shows the numerical estimates as a function of the Bayesian solution. If every participant would have given the Bayesian response, the slope would have been one. Both slopes fitting the participants' estimates are smaller than 1, but the slope in the frequency condition is significantly steeper than in the percentage condition. This is partly due to the fact that the proportion of Bayesian responses was higher in the frequency condition; however, this interaction effect also persists

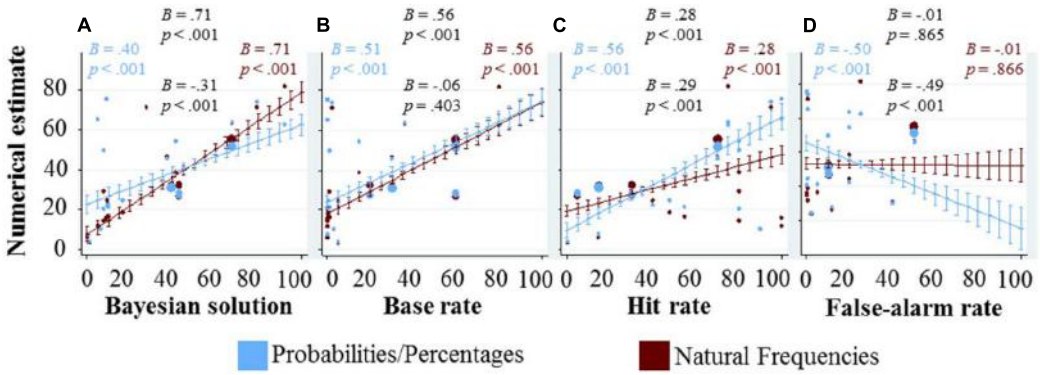

FIGURE 2 | Graphical exploration of how the quantitative variables (displayed at the x-axes) affect the numerical estimates (displayed at the y-axis). Each blue dot represents the average response in the probability/percentage version of a task, each red dot the average response in the natural frequency version, and the size of each dot indicates the number of responses on which these averages are based (see also the rightmost column of **Table 1**). The marginal effects and their confidence intervals are based on regression analysis. In **(A)** showing the Bayesian solution on the *x*-axis, the regression included only representation format and the Bayesian solution as main effects, as well as their interaction. It does not include the three quantitative variables as control variables, as the Bayesian solution already combines them (according to Bayes' rule). The other three panels **(B–D)** display the three quantitative variables, base rate, hit rate, and false-alarm rate on the *x*-axis. Each of the marginal effects and confidence intervals was computed with representation format and all three quantitative variables as main effects as well as interaction effects between representation format and each of the three quantitative variables. In all regressions, standard errors were clustered for each participant. We included four coefficients and corresponding *p*-values in each panel. If at least one of the four *p*-values in a given panel was lower than 0.1, then all four coefficients and their corresponding *p*-values are displayed; otherwise not a single number is reported. The reported numbers are arranged in the shape of a diamond, the main effect on top and the interaction with representation format at the bottom. On the left, we included the main effect when calculating the regression only for responses in the probability/percentage condition, and on the right the main effect in the natural frequency condition. It does not come as a surprise that the numbers on top (main effect) and the numbers on the right (main effect in the natural frequency condition) are almost identical: the interaction is coded as the interaction with the probability/percentage condition, thus the main effect captures the effect in the natural frequency condition. The number on the left can thus also be calculated by adding the main effect (the numbers on top) and the interaction (numbers on the bottom).

when looking only at the non-Bayesian responses ($B = -0.21$, $p < 0.001$). In particular, the slope in the probability/percentage condition decreased from $B = 0.40$ when all responses were taken into account to $B = 0.30$ when considering only the non-Bayesian responses, and the slope in the frequency condition decreased from $B = 0.71$ to $B = 0.51$ (all p's < 0.001). Moreover, we found that participants that were presented information in terms of natural frequencies were more likely to respond with the Bayesian solution; and if they did not, their responses were on average closer to the Bayesian solution [note that this decrease in average absolute differences among the non-Bayesian responses could not be observed for the subset of the four tasks taken from Hoffrage et al. (2015)—to the contrary, there we even found the opposite].

Figure 2B shows that, when statistically controlling for the other two quantitative variables, higher base rates lead to higher numerical estimates. As we have discussed above when introducing the odds version of Bayes' theorem, the prior odds, represented by the base rate, should be positively correlated to the posterior odds. Not surprisingly, such a positive correlation could also be observed between base rates and participants' numerical estimates of posterior probabilities. Again, this effect is partly driven by participants who give the Bayesian response, however, it persists even after all Bayesian responses have been excluded from the regression; in fact, this exclusion reduced the coefficient ($B = 0.31$) but it still remains significant ($p < 0.001$). At the same time, higher base rates are also associated with higher absolute differences between numerical estimates and Bayesian solutions (overall $B = 0.09$, $p = 0.016$; and when only considering the non-Bayesian responses $B = 0.25$, $p < 0.001$). In sum, while more participants find the Bayesian solution for tasks with higher base rates, those participants who do not find the Bayesian solution make larger mistakes in these tasks.

Similarly to the base rate, the hit rate also has a positive effect on the numerical estimate, as can be seen **Figure 2C**. As for the other analyses (**Figures 2A,B**), this effect also persists after all Bayesian responses have been excluded from the regression analysis ($B = 0.35$, $p < 0.001$). Interestingly, in the frequency condition the influence of the hit rate on the numerical estimate is significantly weaker. Additional analyses reveal that only in the probability/percentage condition, higher hit rates are associated with higher absolute deviations from the Bayesian response ($B = 0.30$, $p < 0.001$), and that this effect persists after excluding all Bayesian responses from the analysis ($B = 0.28$, $p < 0.001$).

In **Figure 2D**, it can be observed that the false-alarm rate is strongly negatively related to the numerical estimate in the percentage condition, but unrelated in the frequency condition. The partial correlation between the false-alarm rate and the Bayesian solution (after statistically controlling for the base rate and the hit rate) is -0.13, implying that participants in the probability/percentage condition are overreacting to the false-alarm rate, and participants in the natural frequency condition are not reacting enough. Note that in none of the 19 problems, the false-alarm rate was above 50%, and thus the marginal effects estimates for this area are based on pure extrapolation.

In sum, the three numbers provided in the task are related to the Bayesian solution and Bayes' rule quantifies how the exact relationships are. Generally speaking, the higher the base rate, the higher the Bayesian solution; the higher the hit rate, the higher the Bayesian solution; and the higher the false-alarm rate, the lower the Bayesian solution. Each of these three relationships could be found for the numerical estimates as well. Interestingly, they could also be found even among the non-Bayesian responses. When establishing these relationships for one of the three quantitative dimensions through regression analyses, we controlled for the other two. Note that this statistical control has its limits, because a regression can only do so through a linear combination, while Bayes' rule is not a simple linear combination. In a way, Bayes' rule is the normative correct way how to take all three pieces of information into account, and this is exactly what we have done in **Figure 2A**—which nicely shows that numerical estimates could very well be predicted through the three numbers provided in the task.

How are the Qualitative Variables Related to the Numerical Estimates?

To see how the three qualitative variables characterizing a given task affect the numerical estimates of the participants, we adapted the data representation of our previous question as follows: in **Figure 3**, the three qualitative variables are depicted on the respective *x*-axis of the three panels, and the numerical estimates are plotted on the *y*-axis. As in our previous figure, we again display a dot for the mean numerical estimate of each task in each representation format conditions, and combine this with marginal effects and their confidence interval from regression analysis. The marginal effects of each of the qualitative variables are calculated in a separate regression that only includes the respective qualitative variable, representation format and their interaction. We used separate regressions to explore the differences in responses between tasks, as if the qualitative variable was the only dimension on which the tasks differed. Thus, all the differences between the tasks with a specific quality (e.g., norm deviation = 1) and the tasks without that quality (e.g., norm deviation = 0) will be reflected in the marginal effect shown in **Figure 3**. These marginal effects can thus be seen as an upper bound of the effect of the qualitative variable (unless these qualitative variables are confounded with others factors that have an opposing effect, if this were the case, then 'controlling' for these other factors will increase the observed effects).

Figure 3A shows that the numerical estimates are lower for norm deviation tasks. This negative effect is (significantly) larger within the frequency condition compared to the probability/percentage condition. This pattern can party be explained by the lower Bayesian solutions for the problems where norm deviation is 1, and by the larger number of Bayesian responses in the frequency condition. However, even when excluding all Bayesian responses, both the main effect ($B = -10.88$, $p < 0.001$) and the interaction ($B = 8.99$, $p = 0.007$) remain significant.

Figure 3B depicts a similar, yet less pronounced pattern for the variable stakes. Like for norm deviation, this pattern is partly, but not only, driven by differences in the percentage of Bayesian responses (when only considering the non-Bayesian responses: $B = -10.3$, $p = 0.002$ for the main effect and $B = 8.3$, $p = 0.065$ for the interaction with representation format).

Figure 3C visualizes the effect of the variable main focus. Participants gave significantly higher numerical estimates for tasks in which the main focus was on the individual case, compared to tasks where the main focus was on the numbers. In contrast to the effects depicted in the other two panels,

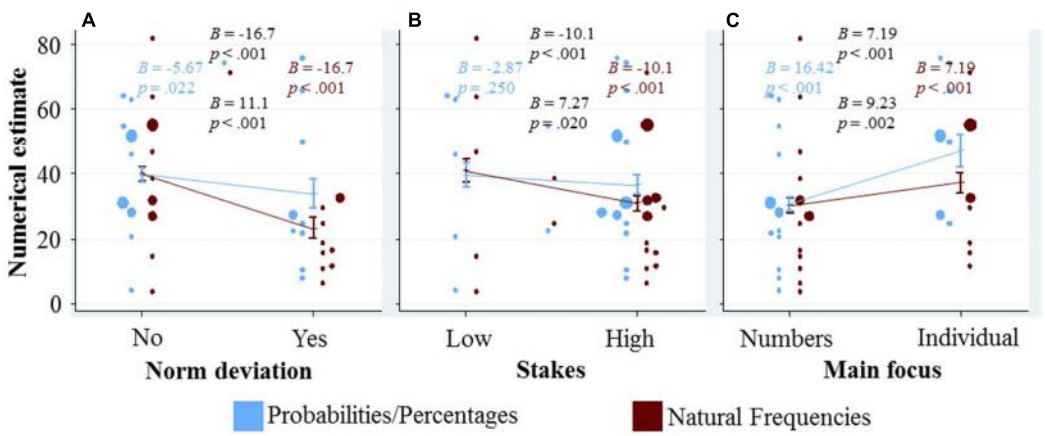

FIGURE 3 | Graphical exploration of how the qualitative variables (displayed at the x-axes) affect the numerical estimates (displayed at the y-axis). To avoid overlap of the dots (see the caption of **Figure 2** for details what they represent), the blue dots are displayed slightly to the left of the confidence interval for the marginal effect, and the red dots slightly to the right. In addition, blue (red) dots that would overlap with other blue (red) dots are moved slightly further to the left (right). The marginal effects and their confidence intervals for each of the qualitative variables **(A–C)** are calculated in a separate regression that only includes the respective qualitative variable, representation format and their interaction (SE were clustered for each participant). We included the four resulting coefficients and corresponding *p*-values in each panel (see caption of **Figure 2** for details).

the effect of main focus is significantly more pronounced in the probability/percentage condition than in the frequency condition. This is particularly interesting, as the Bayesian solutions seem to be unaffected by this variable ($r = 0.05$), and thus a main focus on the individual seems to distract participants from finding the Bayesian solution (and this distraction effect is stronger in the probability/percentage condition).

How does Representation Format Affect the Usage of Cognitive Strategies?

In the previous sections we focused, unless otherwise noted, on all responses and took the numerical estimates as the dependent variable. We will now restrict the analyses only to those responses that have been categorized as Bayesian, or that were identical to either the base rate provided in the task, the hit rate, or the joint occurrence of *D* and *H*. Across both representation formats, any of these four strategies was used in 52.3% of our 1,773 responses. The most frequent strategy, across both formats, was the Bayesian strategy (with 27.7%). The second most often used strategy was the hit rate, but with 11.1% it was used far less often than in other studies (e.g., Villejoubert and Mandel, 2002). The third and fourth most often used strategies were joint occurrence (with 9.2%) and base rate (with 4.3%), respectively.

How did strategy use depend on format? Averaged across all participants and all 19 tasks, the Bayesian strategy was used in 16.9% of cases for probability/percentage representations, and in 38.5% of cases for natural frequency representations ($p < 0.001$, in a logistic regression with standard errors clustered for each participant). For the base rates, these numbers were 6.4 and 2.3%, respectively ($p < 0.001$), and for joint occurrence, 12.1 and 6.2%, respectively ($p < 0.001$). In contrast, format did not exert a significant effect on responding with the hit rate (10.4 vs. 11.7%, respectively; $p = 0.61$) and also not on the usage of the false-alarm rate (1.2 vs. 2%, respectively; $p = 0.23$; the false-alarm rate is not displayed in the Figures and will no longer be considered in the analyses below).

How do the Quantitative Dimensions Affect the Usage of Cognitive Strategies?

As in the last figures, **Figure 4** combines scatter-plots to represent the different tasks in both representation format conditions, with marginal effects and their confidence intervals from regression analysis. In the panels depicted in the first row (**Figures 4A–D**), we explore how the quantitative variables influence participants' performance in finding the Bayesian solution. **Figure 4A** shows that the percentage of participants responding with the Bayesian solutions does not depend on what the Bayesian solution is. In **Figure 4B**, there is a trend indicating that the higher the base rate, the more participants find the Bayesian solution. The effect of the hit rate, depicted in **Figure 4C**, depends on the representation format. In the probability/percentage condition, higher hit rates seem to lead to less Bayesian responses, whereas in the frequency condition, the effect of the hit rate seems to be smaller and in the opposite direction. A potential explanation can be found in the other panels of the third column. When the hit rate is low, participants in the probability/percentage condition used the base rate and the joint occurrence more often as a response strategy. In the last panel of the first row (**Figure 4D**), it can be seen that the higher the false-alarm rate, the smaller the percentage of participants who found the Bayesian solution. In the probability/percentage condition, this can again be partly explained by a higher reliance on the hit rate and joint occurrence as a response strategy. In the frequency condition, however, it is unclear which strategy those participants used who failed to find the Bayesian solution. Note that in the 19 tasks we investigated, the highest false-alarm rate was at 50%, which makes the estimates in the right part of the panel based on pure extrapolation.

How do the Qualitative Dimensions Affect the Usage of Cognitive Strategies?

In **Figure 5**, we explore graphically the effect of the three qualitative variables, norm deviation, stakes and main focus (in the three columns) on the response strategies (in the four rows), again using a combination of scatter-plots and marginal effects with confidence intervals. For the marginal effects and their confidence intervals, we calculated a separate logistic regression for each panel because we did not want to explore the unique contribution of the quantitative variables, but rather the upper bound of their explanatory power, under the assumption that they represent the only difference between the tasks (as in **Figure 3**).

The panels in the first row depict the effects of the qualitative variables on the percentage of Bayesian responses. **Figure 5A** shows that in the probability/percentage condition, it is harder for participants to find the Bayesian solution for tasks with norm deviation, while in the frequency condition the percentage of Bayesian responses did not seem to depend on whether a task includes a norm deviation or not. In **Figure 5B**, it can be seen that whether a task has high or low stakes does not significantly affect the percentage of Bayesian responses. For the sake of completeness, let us mention that when the stakes are high (compared to low), participants in the probability/percentage condition seem to respond more often with the base rate (**Figure 5E**), less often with the hit rate (**Figure 5H**), and more often with the joint occurrence (**Figure 5K**)—while participants in the frequency condition remain largely unaffected by the stakes. A main focus on the individual seems to negatively affect participants' performance in finding the Bayesian solution, especially in the percentage condition (**Figure 5C**). Instead, slightly more participants used the base rate as a response (**Figure 5F**).

Discussion

In this paper we explored the effects of three quantitative and three qualitative dimensions characterizing Bayesian inference tasks on participants' responses. To accomplish this, we plotted the responses of 500 participants to 19 different tasks in several ways. We started broadly with the numerical estimates participants provided as responses, and afterwards classified some of their responses as stemming from different response strategies. We differentiated the tasks both based on the quantitative variables that define the statistical problem—namely,

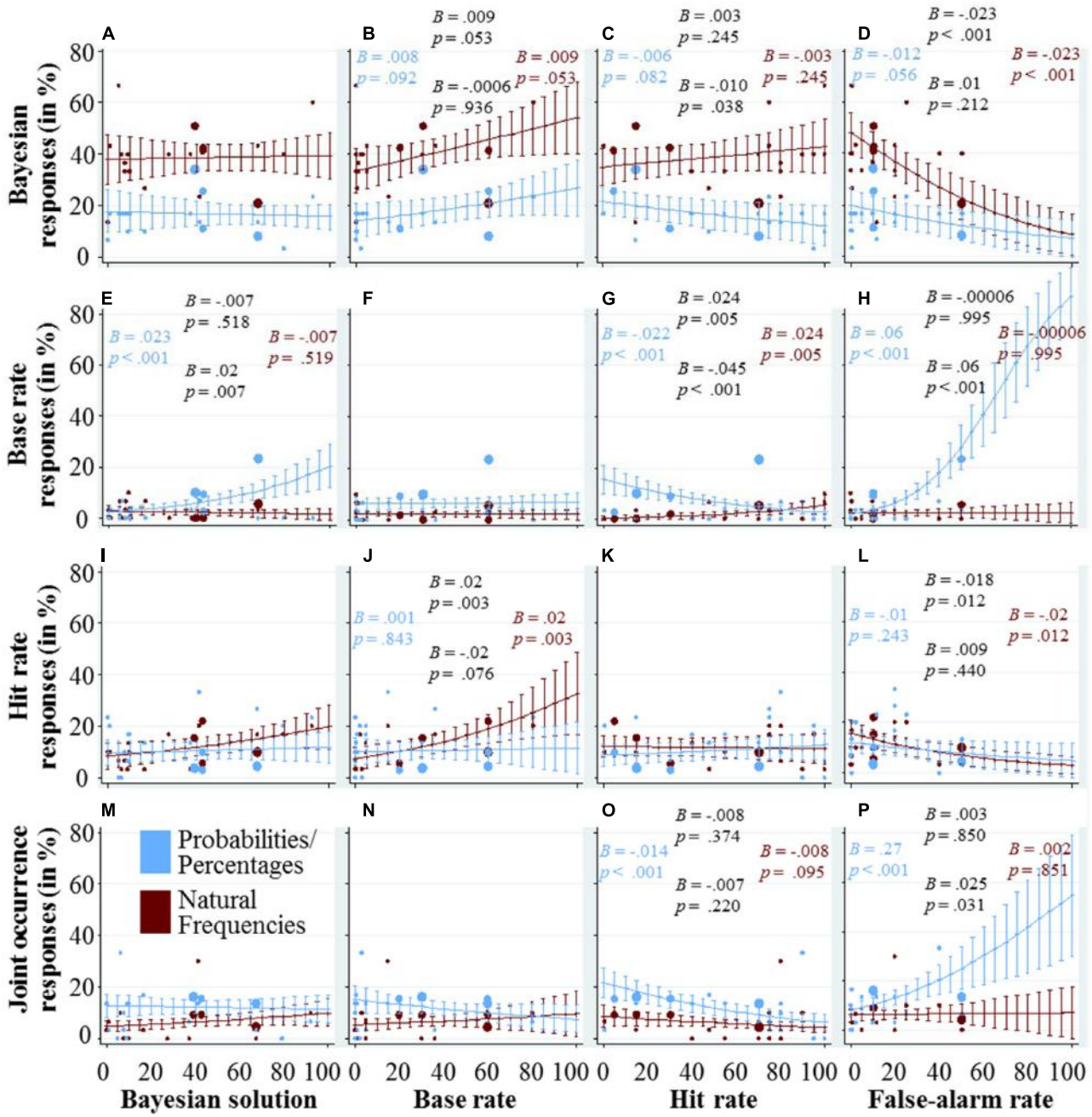

FIGURE 4 | Each of the panels (A–P) visualizes how the quantitative variable that is displayed at the x-axis affects the usage of the cognitive strategy that is displayed at the y-axis. The dots represent the average usage in the different tasks and versions (see caption of **Figure 2** for details). The marginal effects and their confidence intervals are based on logistic regression analysis with strategy use as the dependent variable, following the same specifications as the regression analyses used for **Figure 2** (see the caption of **Figure 2** for details). We also included the four coefficients and corresponding *p*-values in each panel (again, see the caption of **Figure 2** for details).

the base rate, hit rate, and false-alarm rate—and on qualitative variables that describe the context and narrative of the task. In this explorative analysis, we found that participants seem not to perceive all Bayesian inference tasks as being equal, and most of the variables we investigated seem to influence not only the specific numeric response participants are providing, but—and of course not independently of the numeric responses—which strategy they use. In the remainder of this paper, we want to highlight three main lessons we draw from this exploratory investigation, and we outline some avenues for future research.

First, the numerical value of the Bayesian solution does not seem to influence whether participants find it. While their

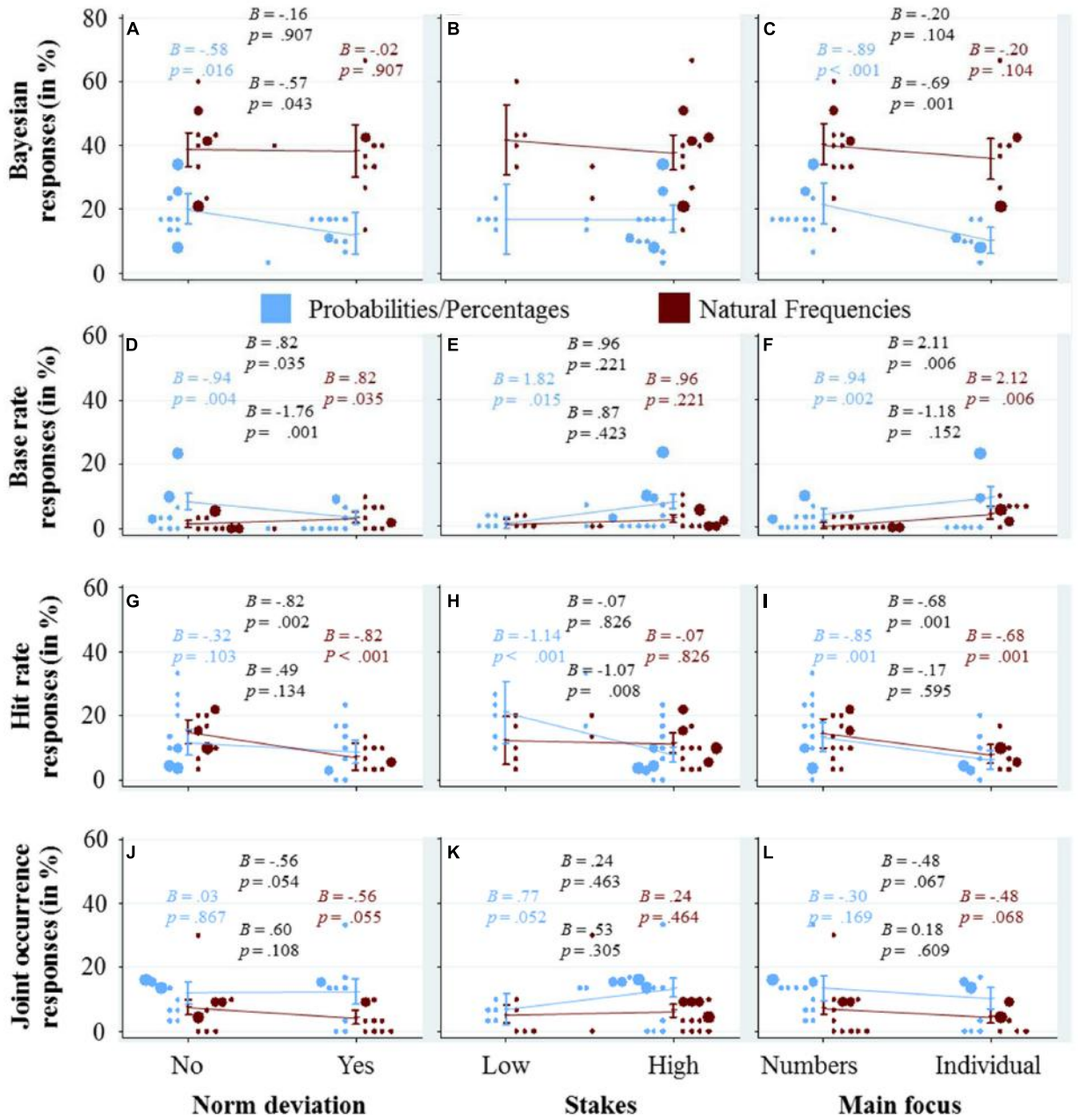

FIGURE 5 | Each of the panels (A–L) visualizes how the qualitative variable that is displayed at the x-axis affects the usage of the cognitive strategy that is displayed at the y-axis. The dots represent the average usage in the different tasks and versions (see caption of **Figure 2** for details about their color and size). The marginal effects and their confidence intervals are based on logistic regression analysis with strategy use as the dependent variable, following the same specifications as the regression analyses used for **Figure 3** (see the caption of this figure for details). We included four coefficients and corresponding *p*-values in each panel (see also the caption of **Figure 2** for details).

responses are driven by the different pieces of information stated in the task (base rate, hit rate, and false-alarm rate), and by other qualitative variables that can be seen as irrelevant from a normative point of view, their response strategy seems to be unaffected by what the Bayesian solution is. For our set of 19 tasks, about the same proportion of individuals provides the Bayesian solution, independent of whether it is as low as 0.03% or as high as 92.3%. However, the large majority of the participants' responses (77.9%) were from a task for which the Bayesian solution was 42.9% or less.

Second, focusing on the numbers, instead of the individual case, seems to increase participants' performance. Interestingly, this effect was more pronounced in the probability/percentage

condition and less pronounced in the natural frequency condition. To better understand the effect of main focus, it is useful to consider the debate about the underlying mechanism of the beneficial effect of natural frequencies (Gigerenzer and Hoffrage, 2007; Brase, 2008; Hill and Brase, 2012; Brase and Hill, 2015; Johnson and Tubau, 2015). The two prominent explanations for the beneficial effect of natural frequencies are that they (a) make the nested set relationship more explicit, and that they (b) prompt participants to think in terms of frequencies (instead of "single event probabilities"). Brase (2008) provided evidence for the second explanation: participants who interpreted the somewhat ambiguous word "chances" as frequencies performed better than those participants who interpreted "chances" as probabilities. In line with this explanation, a main focus on the numbers might lead participants to adopt a frequentist point of view, thereby increasing their performance. In contrast, a main focus on the individual case might prevent participants from adopting and, in turn, benefiting from such a viewpoint. This account would also explain why the effect of main focus was more pronounced for probability representations, where Bayesian performance tends to be low and thus leaves more room for the effect of focusing on the numbers—while for natural frequency representations a main focus on the numbers had less added value as most participants were already thinking in terms of numbers anyway (but the effect of main focus could still be observed even within the frequency condition, see **Figure 5C**).

The practical consequences of the main focus might be particularly severe, as the tasks with a focus on the individual case tend to be about a norm deviation and tend to have high stakes, at least in our sample of tasks. Of course we cannot make any causal claims here, but our results are consistent with the following speculation: if a specific problem involves a norm deviation and if stakes are high, those who formulate a problem may be led to focus on the individual case, for instance, to attract the readers' attention, to appeal to emotions, and to increase empathy (cf. the identified-victim effect, Small and Loewenstein, 2003). They may even adopt such a focus with good intentions, namely to increase the readers' involvement and motivation to solve the problem. And even if a task description is relatively neutral, chances are that the reader may focus on the individual if the hypothesis involves a norm deviation and if stakes are high. However, and ironically, such a frame increases the difficulty of the problem, as our results suggest, and may more than offset any beneficial effect that the increased motivation and the personal affection might have. It may sound trivial, but this points to a potential strategy how problems could be reframed (or how individuals could reframe them in their head) to boost the accuracy of responses: use natural frequencies rather than probabilities to communicate the statistical information, and, on top of this, focus on the numbers rather than on the individual case. However, as our analysis is only exploratory, future research would be needed to systematically test such a reframing strategy, and to disentangle the effect of 'main focus' of the task from other effects and to identify potential boundary conditions.

Third, in the probability/percentage condition, the quantitative and qualitative task characteristics influenced participants' responses to a larger extent than in the natural frequency condition. This could possibly be explained by the fact that the percentage of Bayesian responses was higher in the natural frequency condition (on average there were 38.5% Bayesian responses in the natural frequency condition and only 16.9% in the probability/percentage condition). For someone who figured out how to structurally solve the Bayesian inference tasks (Johnson and Tubau, 2015), there was no need to find a solution in the particulars of the task specific context stories or to use a non-Bayesian strategy, for instance, by taking one of the numbers provided in the task or by integrating them in some other way. In contrast, someone who did not understand how the numbers should be combined could be tempted to look for similarities between the problem at hand, and problems they have solved before. In other words, for someone who figured out what the normative response strategy is, the task content and any other characteristics were exchangeable decoration—and for those who did not, such variables could possibly exert an influence.

Yet, in the natural frequencies condition, many participants also struggled with the tasks and where hence vulnerable to task dimensions that are irrelevant from a normative point of view. Why are Bayesian tasks still hard for some, even when information is presented in terms of natural frequencies? One reason could be that outside of the lab, most situations in which individuals update their beliefs do not feature numerical information about base rates, hit rates, and false-alarm rates. For some participants, it might have been the first time that they encountered such text book problems when they read the descriptions of the tasks in the context of the experiment. Outside of the lab, information updating might often rather consist of evaluating some data/sampling some information, based on which individuals form their initial beliefs, and then afterward evaluating some more data, maybe more locally relevant or more recent, and then revising their beliefs in light on the new data. Of course, in their statistical structure, such situations are different from Bayesian inference tasks, but because individuals have much more experience with other information updating tasks, they might try to rely on this experience to make sense of the Bayesian inference tasks. And, outside of the environment of Bayesian inference tasks, information updating strategies that are contingent on task specific factors such as the trustworthiness of the initial data or the new data (Welsh and Navarro, 2012) or the judgment of the validity of the sample (Fiedler, 2000) might be ecologically rational.

Overall, we can conclude from these exploratory analyses that not only the quantitative variables (the numbers given in the task) but also our qualitative variables (norm deviation, stakes and main focus) could explain some variance in participants' responses and in particular in the strategies they use. However, even though most of the effects of our six predictor variables on the four strategies that we inspected reached statistical significance, we hasten to add that such a result is not too hard to achieve with 1,773 responses and that most of the differences

between the percentage points of strategy use, contingent on the levels of our dichotomous predictor variables, was in the order of five percentage points. Given that most percentages were close to the lower end of the scale, such differences are, relatively speaking, quite large, but with respect to the whole scale of 100%, a difference of 5% points is still a small difference. Moreover, it is important to consider that this analysis is based on a *post hoc* analysis of only 19 tasks, and that these tasks were not designed to allow for systematic tests of the quantitative and qualitative dimensions. For instance, as norm deviation is highly correlated with the base rate and the Bayesian solution (and not orthogonally manipulated), it is not possible to causally attribute the observed effect to either the qualitative dimension (i.e., the norm deviation narrative) or the underlying quantitative dimensions (i.e., the numbers).

One avenue for future research could hence be to use constructed scenarios and manipulate some of the variables used in the present analysis systematically, that is, orthogonally. A prime example for this approach is Krynski and Tenenbaum's (2007) study that we mentioned in Section "Introduction": these authors manipulated one aspect of the task while keeping everything else constant. This would naturally allow for conclusions that have a much higher internal validity, compared to the observations we can share and the tentative conclusions we can formulate based on our exploratory analyses, which were based on a comparison between tasks that differed on many aspects simultaneously.

Another avenue would be to go in the opposite direction: not to use systematic designs, but what Brunswik called a representative design (see Dhami et al., 2004). Even though we referred to the present analyses as a first step toward an ecological analysis of Bayesian inferences, we must acknowledge that it does not fully deserve this label. For many of the 19 tasks, the base rates and the statistical properties of the diagnostic test have been made up rather than measured in a real-world context. It would hence be interesting to conduct such an analysis and to study the dimensions that may affect strategy use in larger pool of Bayesian inference tasks from real-world applications and with natural inter-correlations between the variables of interest.

For many study participants, Bayesian inference tasks are hard, and most responses are not Bayesian. Moreover, the qualitative task characteristics that we scrutinized in our analyses should not play a role from a normative point of view, however, they did influence participants' responses and they also had an impact on which cognitive strategy they used. How can one account for non-normative responses and for the finding that task characteristics that should be irrelevant from a normative point of view *did* play a role? A promising approach to answer this question may involve making an attempt to put oneself into participants' shoes and to ask how they approach the task. Which mental models (Gentner and Stevens, 1983; Johnson-Laird, 1983) do they construct? What is their problem space (Simon and Newell, 1971; Newell and Simon, 1972)? What kinds of belief updating tasks do they encounter in their environments and how could their experience with these tasks possibly inform solutions to this special class of belief updating tasks that come in the form of textbook problem? As researchers who study how participants change their beliefs in light of new data, we may eventually find out that we may need to change our perspective, research questions, and research paradigms in light of new experimental findings. Adopting the perspective of individuals who have to solve Bayesian tasks, and aiming at understanding what constitutes the environment of comparable tasks from their perspective seems to be a fruitful avenue for future research.

Acknowledgments

We would like to thank Guillaume Blanc, Justin Olds, Tim Pleskac, Jan K. Woike and the two reviewers and research topic editors for helpful comments, and Matthieu Legeret for assisting us with producing the figures. This work was supported by grant 100014_140503 from the Swiss National Science Foundation.

References

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychon. Bull. Rev.* 15, 284–289. doi: 10.3758/PBR.15.2.284

Brase, G. L., and Hill, W. T. (2015). Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Front. Psychol.* 6:340. doi: 10.3389/fpsyg.2015.00340

Cosmides, L. (1989). The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition* 31, 187–276. doi: 10.1016/0010-0277(89)90023-1

Dhami, M., Hertwig, R., and Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychol. Bull.* 130, 959–988. doi: 10.1037/0033-2909.130.6.959

Fiedler, K. (2000). Beware of samples! a cognitive-ecological sampling approach to judgment biases. *Psychol. Rev.* 107, 659–676. doi: 10.1037/0033-295X.107.4.659

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

Gentner, D., and Stevens, A. L. (eds). (1983). *Mental Models*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: a reply to Kahneman and Tversky (1996). *Psychol. Rev.* 103, 592–596. doi: 10.1037/0033-295X.103.3.592

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., and Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev.* 106, 425–430. doi: 10.1037/0033-295X.106.2.425

Gigerenzer, G., and Hoffrage, U. (2007). The role of representation in Bayesian reasoning: correcting common misconceptions. *Behav. Brain Sci.* 30, 264–267. doi: 10.1017/S0140525X07001756

Griffiths, T. L., and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychol. Sci.* 17, 767–773. doi: 10.1111/j.1467-9280.2006.01780.x

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychol. Sci.* 15, 534–539. doi: 10.1111/j.0956-7976.2004.00715.x

Hill, W. T., and Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *Q. J. Exp. Psychol.* 65, 2343–2368. doi: 10.1080/17470218.2012.687004

Hoffrage, U., and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Acad. Med.* 73, 538–540. doi: 10.1097/00001888-199805000-00024

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition* 84, 343–352. doi: 10.1016/S0010-0277(02)00050-1

Hoffrage, U., Hafenbrädl, S., and Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Front. Psychol.* 6:642. doi: 10.3389/fpsyg.2015.00642

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science* 290, 2261–2262. doi: 10.1126/science.290.5500.2261

Johnson, E. D., and Tubau, E. (2015). Computation and comprehension in Bayesian problem solving. *Front. Psychol.* 6:938. doi: 10.3389/fpsyg.2015.00938

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* Cambridge, MA: Harvard University Press.

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Koehler, J. J. (1996). The base rate fallacy reconsidered: descriptive, normative and methodological challenges. *Behav. Brain Sci.* 19, 1–53. doi: 10.1017/S0140525X00041157

Krynski, T. R., and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *J. Exp. Psychol. Gen.* 136, 430–50. doi: 10.1037/0096-3445.136.3.43

Macchi, L. (1995). Pragmatic aspects of the base rate fallacy. *Q. J. Exp. Psychol.* 48A, 188–207. doi: 10.1080/14640749508401384

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Front. Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Front. Psychol.* 6:387. doi: 10.3389/fpsyg.2015.00387

McNair, J. (2015). Beyond the status-quo: research on Bayesian reasoning must develop in both theory and method. *Front. Psychol.* 6:97. doi: 10.3389/fpsyg.2015.00097

Mellers, B. A., and McGraw, A. P. (1999). How to improve Bayesian reasoning: comment on Gigerenzer and Hoffrage (1995). *Psychol. Rev.* 106, 417–424. doi: 10.1037/0033-295X.106.2.417

Newell, A., and Simon, H. A. (1972). *Human Problem Solving*, Vol. 104, No. 9. Englewood Cliffs, NJ: Prentice-Hall.

Pleskac, T., and Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *J. Exp. Psychol. Gen.* 143, 2000–2019. doi: 10.1037/xge0000013

Simon, H. A., and Newell, A. (1971). Human problem solving: the state of the theory in 1970. *Am. Psychol.* 26, 145. doi: 10.1037/h0030806

Sirota, M., Vallée-Tourangeau, G., Vallée-Tourangeau, F., and Juanchich, M. (2015). On Bayesian problem-solving: helping bayesians solve simple bayesian word problems. *Front. Psychol.* 6:1141. doi: 10.3389/fpsyg.2015.01141

Small, D. A., and Loewenstein, G. (2003). Helping a victim or helping the victim: altruism and identifiability. *J. Risk Uncertain.* 26, 5–16. doi: 10.1023/A:1022299422219

Todd, P. M., Gigerenzer, G., and the ABC Research Group. (2012). *Ecological Rationality: Intelligence in the World*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195315448.003.0011

Villejoubert, G., and Mandel, D. R. (2002). The inverse fallacy: an account of deviations from Bayes's theorem and the additivity principle. *Mem. Cogn.* 30, 171–178. doi: 10.3758/BF03195278

Welsh, M. B., and Navarro, D. J. (2012). Seeing is believing: priors, trust, and base rate neglect. *Organ. Behav. Hum. Decis. Process.* 119, 1–14. doi: 10.1016/j.obhdp.2012.04.001

Wolfe, C. R. (1995). Information seeking on Bayesian conditional probability problems: a fuzzy-trace theory account. *J. Behav. Decis. Mak.* 8, 85–108. doi: 10.1002/bdm.3960080203

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 17 August 2015
doi: 10.3389/fpsyg.2015.01194

Bayesian probability estimates are not necessary to make choices satisfying Bayes' rule in elementary situations

Artur Domurat[1], Olga Kowalczuk[2], Katarzyna Idzikowska[3], Zuzanna Borzymowska[4] and Marta Nowak-Przygodzka[1]*

[1] Department of Cognitive Psychology, Faculty of Psychology, University of Warsaw, Warsaw, Poland, [2] Center for Complex Systems and New Technologies, The Robert B. Zajonc Institute for Social Studies, University of Warsaw, Warsaw, Poland, [3] Centre for Economic Psychology and Decision Sciences, Kozminski University, Warsaw, Poland, [4] Laboratory of Visual System, Nencki Institute of Experimental Biology, Warsaw, Poland

OPEN ACCESS

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
Gaëlle Vallée-Tourangeau, Kingston University London, UK
David E. Over, Durham University, UK

****Correspondence:***
Artur Domurat, Department of Cognitive Psychology, Faculty of Psychology, University of Warsaw, Stawki 5/7, Warsaw 00-183, Poland
artur.domurat@psych.uw.edu.pl

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *27 February 2015*
Accepted: *28 July 2015*
Published: *17 August 2015*

Citation:
Domurat A, Kowalczuk O, Idzikowska K, Borzymowska Z and Nowak-Przygodzka M (2015) Bayesian probability estimates are not necessary to make choices satisfying Bayes' rule in elementary situations. Front. Psychol. 6:1194. doi: 10.3389/fpsyg.2015.01194

This paper has two aims. First, we investigate how often people make choices conforming to Bayes' rule when natural sampling is applied. Second, we show that using Bayes' rule is not necessary to make choices satisfying Bayes' rule. Simpler methods, even fallacious heuristics, might prescribe correct choices reasonably often under specific circumstances. We considered elementary situations with binary sets of hypotheses and data. We adopted an ecological approach and prepared two-stage computer tasks resembling natural sampling. Probabilistic relations were inferred from a set of pictures, followed by a choice which was made to maximize the chance of a preferred outcome. Use of Bayes' rule was deduced indirectly from choices. Study 1 used a stratified sample of $N = 60$ participants equally distributed with regard to gender and type of education (humanities vs. pure sciences). Choices satisfying Bayes' rule were dominant. To investigate ways of making choices more directly, we replicated Study 1, adding a task with a verbal report. In Study 2 ($N = 76$) choices conforming to Bayes' rule dominated again. However, the verbal reports revealed use of a new, non-inverse rule, which always renders correct choices, but is easier than Bayes' rule to apply. It does not require inversion of conditions [transforming P(H) and P(D|H) into P(H|D)] when computing chances. Study 3 examined the efficiency of three fallacious heuristics (pre-Bayesian, representativeness, and evidence-only) in producing choices concordant with Bayes' rule. Computer-simulated scenarios revealed that the heuristics produced correct choices reasonably often under specific base rates and likelihood ratios. Summing up we conclude that natural sampling results in most choices conforming to Bayes' rule. However, people tend to replace Bayes' rule with simpler methods, and even use of fallacious heuristics may be satisfactorily efficient.

Keywords: Bayes' rule, choices, binary hypothesis, heuristics, natural sampling, ecological rationality, non-inverse rule

Introduction

This paper aims to investigate whether people conform to Bayes' rule when making choices in probabilistic situations, or whether they tend to simplify their reasoning by using other methods. To develop an understanding of what a Bayesian problem is, consider the following example:

The red nose problem (Zhu and Gigerenzer, 2006, p. 289).

Pingping goes to a small village to ask for directions. In this village, 10 out of every 100 people will lie. Of the 10 people who lie, eight have a red nose. Of the remaining 90 people who do not lie, nine also have a red nose. Imagine that Pingping meets a group of people in the village with a red nose. How many of these people will lie?

The red nose problem illustrates an elementary situation, which is defined with binary sets of hypotheses (H and not-H) and data (D and not-D). A person can lie or not and have a red or non-red nose. There can be several hypotheses and data sets, but discussing such situations is beyond scope of the present article. According to Bayes' rule, an estimate of the posterior probability of a distinct hypothesis should be computed using the observations provided and the prior probability of the hypothesis. In the example, the goal is to compute the posterior probability of being a liar given that a person has a red nose. We denote it with P(H|D) and compute it with the formula:

$$P(H|D) = P(H \text{ and } D) \,/\, P(D) = \quad (1)$$
$$P(H)P(D|H) \,/\, [\, P(H)\, P(D|H) \,+\, P(\text{not-H})\, P(D|\text{not-H})]$$

To perform the calculation we need to know the base rate P(H), which is the chance of a person being a liar, P(H) = 10/100 = 10%. Thus, the chance of being a non-liar, P(not-H) is 90%. This should be updated with new data, conditional probabilities P(D|H) and P(D|not-H). Hence, there are P(D|H) = 8/10 = 80% of people with a red nose among liars, and P(D|not-H) = 9/90 = 10% among non-liars (people who have a red nose when they tell the truth). This enables computation of whether a person with a red nose will lie:

$$P(H|D) = 10\% \times 80\% \,/\, [10\% \times 80\% + 90\% \times 10\%] = 47\%$$

The conclusion is that those with a red nose will lie with a 47% probability.

Bayesian estimates are counter-intuitive and people are usually surprised by the discrepancy among a base rate (10% in the above example), likelihood ratio (80%), and actual Bayesian probability (47%). Similar discrepancies occur in such well-known cases as the taxi cab problem (Tversky and Kahneman, 1982) and the mammography problem (Eddy, 1982). In the taxi cab problem, given a witness's evidence that a cab was blue, the probability that the cab was actually blue is 41%. This Bayesian result differs from – specified in the case – the base rate of 15% and the likelihood ratio of proper color identification which equals 80%. In the mammography problem, while the base rate of breast cancer is 1% and the likelihood ratios are 80% for a positive test and 9.6% for a false alarm, the actual Bayesian probability is 7.8%.

Numerous studies show that people have difficulty in finding solutions for Bayesian problems. Subjects acquainted with new evidence are conservative and underestimate posterior chances (Phillips and Edwards, 1966; Edwards, 1968). They also demonstrate the base rate fallacy, neglecting P(H), and the inverse fallacy, confusing likelihood ratios P(D|H) with Bayesian estimates P(H|D) (Koehler, 1996; Villejoubert and Mandel, 2002). Systematic ignorance of prior probabilities and overuse of the representativeness heuristic have led to the conclusion that people are not Bayesians (Kahneman and Tversky, 1972, 1973; Tversky and Kahneman, 1982).

Misapprehension of the probabilities may lead to inadequate decisions and entail severe consequences. Gigerenzer et al. (1998) reported the case of seven out of 22 blood donors who committed suicide after they were shown to be HIV-positive by the ELISA and Western Blot tests, which had a 100% detection efficiency. It transpired that the actual Bayesian probabilities were around 50%. The authors concluded that there is a need to develop tools for understanding and appropriately communicating risks in AIDS counseling centers. Such problems occur not only in the domain of medical diagnosis but in other domains where probabilistic evaluations depend on both prior distributions and newly obtained information (e.g., in management, law and intelligence analysis – see Nance and Morris, 2005; Hoffrage et al., 2015; Mandel, 2015). A vast amount of research was focused on pedagogical issues surrounding Bayesian inference. Methods were elaborated to aid the understanding of Bayes' rule and facilitate communication of risk appropriately. These used visual representations such as Venn diagrams, trees, pictorial representations, or frequency grids (Mellers and McGraw, 1999; Yamagishi, 2003; Brase, 2008; Mandel, 2014; Navarrete et al., 2014; Sirota et al., 2014).

Bayesian reasoning issues have been of particular interest to evolutionary psychologists, who have proposed an ecological rationality framework for research (Gigerenzer and Hoffrage, 1995; Cosmides and Tooby, 1996; Brase et al., 1998). According to this approach, people are not evolutionarily prepared for performing abstract computations. In particular, the concept of probability is an ecologically invalid notion. The calculus of probability is a relatively recent discovery in humankind's history, and the human mind having evolved to maintain information in the form of absolute numbers. Such numbers are termed natural frequencies and the process of gathering information on natural frequencies through real life experience is termed natural sampling (Kleiter, 1994; Gigerenzer and Hoffrage, 1995; Gigerenzer, 1998). Because humans have collected information in the form of natural frequencies throughout evolution, such representations facilitate correct Bayesian reasoning (Cosmides and Tooby, 1996; Sedlmeier and Gigerenzer, 2001).

For example, the natural frequencies in the red nose problem are:

- The total number of village inhabitants, $a = 100$,

- Numbers of liars, $b = 10$,
- Non-liars, $c = 90$,
- People with a red nose among liars, $d = 8$,
- Liars with no red nose, $e = 2$,
- Non-liars with a red nose, $f = 9$,
- Non-liars with no red nose, $g = 81$.

Studies by Zhu and Gigerenzer (2006) showed that even children can give appropriate answers to Bayesian problems if they are presented with natural frequencies. The frequencies simplify computations because posterior probabilities can be estimated as:

$$\mathrm{P(H|D)} \;=\; d/[\,d+f\,]$$

To compute whether a person with a red nose will lie it is sufficient to calculate:

$$\mathrm{P(H|D)} \;=\; 8/(8+9) \;=\; 47\%$$

The evolutionary approach has been criticized for being difficult to falsify (Girotto and Gonzalez, 2001). While people deal with natural formats better than with probabilities, this does not necessarily mean that this ability has developed through natural selection or adaptation. One cannot simply rely on previous experience to perform successfully in a novel or complex environment. Frequencies help visualize nested sets and relations, and thereby facilitate solution of Bayesian problems, but this does not necessarily result from Bayesian inference (Sloman et al., 2003). Solving probabilistic problems requires also the comprehension of elementary logic, set operations and relations (Barbey and Sloman, 2007). For instance in Girotto and Gonzalez (2001) studies subjects performed better when subset relations were activated.

We agree that the evolutionary approach is not convincing in its explanation of how reasoning developed, and the issue of how the ability to collect and process natural frequencies developed in humans is debatable. However, there is agreement that natural frequencies are easier to process and that people learn about statistical relationships from natural sampling in real life. Hence, the ecological framework seems to be valid at least in that:

(1) Statistical information is gathered via natural sampling,
(2) The environment defines objectives and supplies means to achieve them, and
(3) Human rationality is ecological.

Nevertheless, these propositions lead us to conclude that single probability judgments do not provide sufficient information for attaining goals in situations such as the red nose problem, where choices are placed before people in fact. In the original story (Zhu and Gigerenzer, 2006), Pingping's goal is to obtain the right directions to continue his journey, and he is expected to assess the chance of being cheated by people with a red nose. However, exploring the truthfulness of people with red noses only is not enough: Pingping has to decide whether to ask for directions someone with a red nose or refrain from this and ask a person without a red nose, actually. 'Having no red nose' is also a clue with some ecological validity. Thus, we propose a modified question in the red nose problem:

Should Pingping ask a person with a red nose for directions, or find a person who does not have a red nose?

What works for evaluating the truthfulness of people with red noses will also work for evaluating the truthfulness of people without red noses. To answer the question, Pingping should calculate the proportion of liars among both people with red-noses and people without red noses, applying Bayes' rule twice:

- $\mathrm{P(H|D)} \;=\; d/(d+f) \;=\; 8/(8+9) \;=\; 47\%$
- $\mathrm{P(H|not\text{-}D)} \;=\; e/(e+g) \;=\; 2/(2+81) \;=\; 2\%$

Having compared these chances, Pingping should conclude that he takes a far greater risk of being lied to when he asks someone with a red nose and conclude that it is better to find someone without a red nose.

Reconsidering the red nose problem in such a way shows that, to solve such problems, estimates referring to all the options are needed. This is in the line with probabilistic functionalism, which proposes that people do not evaluate probabilities for their own sake, but to achieve specific goals. People infer missing data from probabilistic indicators to reduce incompleteness and uncertainty in their knowledge (Brunswik, 1943; Dhami et al., 2004; Pleskac and Hertwig, 2014).

There is common agreement that natural sampling may facilitate correct Bayesian reasoning. People acquire knowledge about probabilities from their own experience rather than compiled frequency statistics (Gigerenzer, 1998). Surprisingly, natural sampling is not reflected in most experiments, where participants are provided with well-prepared and well-arranged natural frequencies or probabilities (Kleiter, 1994; Girotto and Gonzalez, 2001). We postulate that experiments should attempt to approximate the experiential aspect of natural sampling. However, such experiments should not give clues to participants about processing data at the same time. An understanding of conditions in general is a crucial step in solving a Bayesian problem. After realizing that the inferential process should be narrowed to a given condition (the first step in Eq. 1), one should invert one's thinking about conditions from $D|H$ into $H|D$ (the second part of Eq. 1). Framing tasks with natural frequencies ("Imagine that Pingping meets a group of people in the village with red noses. How many of these people will lie? __ out of __," as originally in Zhu and Gigerenzer, 2006, p. 289) is suggestive and entails scaffolding the answers. The group characterized by data D is identified directly ("these people") and the subsequent question suggests narrowing thinking to this set ("__ out of __"). A person has no need to perform the first step on their own in tasks framed this way, and the clue about how to answer helps people to avoid comitting the inverse fallacy. Hence, we postulate that research techniques should reflect natural sampling, but in a way that gives no clues to participants about how to process probabilistic information.

In our studies, we mimic the process of natural sampling and present participants with actual events instead of probabilities or frequencies. We anticipate that participants should have learned these from their own experience and that they should make

choices based upon them. This approach reflects a paradigm in which decisions based upon participants' own experience are explored, as proposed by Hertwig et al. (2004) and continued by their followers (for a review, see Hertwig and Erev, 2009; Rakow and Newell, 2010). As these researchers argue, people make everyday decisions, such as backing-up a hard drive or crossing a busy street, by relying on the recall of events that they have previously experienced, not based upon descriptions of outcomes or likelihoods (Hertwig et al., 2004). Everyday decisions or choices rarely need to be articulated in exact numbers and the outcome of one's inference is usually expressed in his actions or choices, not in estimates of probabilities. Therefore, it should be easier for people to deal with Bayesian problems by choosing between two alternatives (differing with respect to a posteriori probabilities of success) rather than giving exact numbers. Hence the first question regarding choices in elementary situations that we aim to answer is:

[Q1] How often do people make choices satisfying Bayes' rule, when probability information is gathered through natural sampling?

In answering Q1, one can expect that [H] choices will conform to Bayes' rule in natural sampling settings:

[Ha] in most of the tasks (a strong criterion) or

[Hb] more frequently than at random (a weak criterion).

On the one hand, people tend to maximize their performance. It should also be easier for people to articulate a solution by choosing between two alternatives, rather than articulating exact numbers. Hence, their choices should comply with Bayesian rule (Ha). On the other hand, using the rule is cognitively costly, so it may often be ignored or replaced with heuristics or other methods. For instance, comparing fractions may turn out to be just as hard as comparing probabilities or percentages (Kleiter, 1994; Gigerenzer and Hoffrage, 1995; Gigerenzer, 1998). Even if fractions are estimated properly, computational complexity increases with the necessity of performing two correct Bayesian evaluations and performing a correct comparison of them when making choices. We therefore also formulated a weaker expectation that the choices would comply with Bayes' rule more frequently than other methods (Hb).

To answer question Q1, we created experiments that reflected natural sampling, with the intention of showing how often people make choices satisfying Bayes' rule (Studies 1 and 2).

Using Bayes' rule requires cognitive effort and only pays-off when one can make significantly better decisions. Cognitive limitations and the avoidance of effort make people turn to the use of fallacious heuristics, which are popular because they are frugal and still roughly correct (Gigerenzer et al., 1999; Gigerenzer, 2004, 2008). As Simon (1955, 1956) hypothesized, people select strategies that meet minimal standards and aspirations. Ecological rationality postulates that calculations do not have to be correct, however, they should be correct reasonably often (Gigerenzer, 2004; Over, 2004). As Gigerenzer (2008, p. 25) further explained, "The goal of an organism is not to follow logic, but to pursue objectives in its environment, such as establishing alliances, finding a mate, and protecting offspring. Logic may or may not be of help. The rationality of the adaptive toolbox is not logical, but *ecological*; it is defined by correspondence rather than coherence."

Summing up, the interesting issue is whether heuristics can prescribe correct answers satisfactorily often, given some specific circumstances. Thus, we raise the question:

[Q2] How often do fallacious heuristics yield choices that conform to Bayes' rule?

Zhu and Gigerenzer (2006) observed that, instead of Bayes' rule, people use the following fallacious heuristics (following these authors, we apply the term "cognitive strategies" or in short "strategies" describing them and Bayes' rule):

- the conservatism strategy: b/a,
- the evidence-only strategy: $(d+f)/a$,
- the representativeness strategy: d/b,
- the pre-Bayesian strategy: $b/(d+f)$.

By analogy, people may apply these cognitive strategies to simplify their choices in elementary situations through the following comparisons:

- the evidence-only strategy: comparing $(d+f)$/a with $(e+g)/a$,
- the representativeness strategy: d/b with e/b,
- the pre-Bayesian strategy: $b/(d+f)$ with $b/(e+g)$, and
- the conservatism strategy: b/a with c/a.

In the red nose problem, the Bayes' rule, the representativeness strategy ($d/b = 8/10 > e/b = 2/10$) and the pre-Bayesian strategy [$10/(8+9) > 10/(2+81)$] would result in a decision not to ask a person with a red nose. Only the evidence-only strategy would render a different conclusion [$(8+9)/100 < (2+81)/100$].

To answer question Q2 we investigated how often fallacious strategies (representativeness, pre-Bayes, and evidence-only) prescribe the same choices as Bayes' rule by carrying out computer simulations of natural frequencies (Study 3).

Study 1

The goal of Study 1 was to answer Q1: how often do choices conform to Bayes' rule in elementary situations?

Materials and Methods

We used a computer program with a sequence of 16 simulation tasks, which we called "adventures." Introductory instructions were as follows: "The study you will be taking part in is aimed at finding out how people find precious objects. You will be presented with 16 opportunities to acquire precious objects: diamonds and amber. Each of the 16 adventures consists of two stages. The initial phase should familiarize you with the area. The second part requires you to identify where the gem is hidden. Each adventure is independent and concerns treasures in the form of diamonds or amber. The next screen will reveal the first phase of adventure number one. You will be presented with seven cards. On the face of each card you will find a diamond (a piece of amber) or a stone (a piece of broken glass). Clicking the card

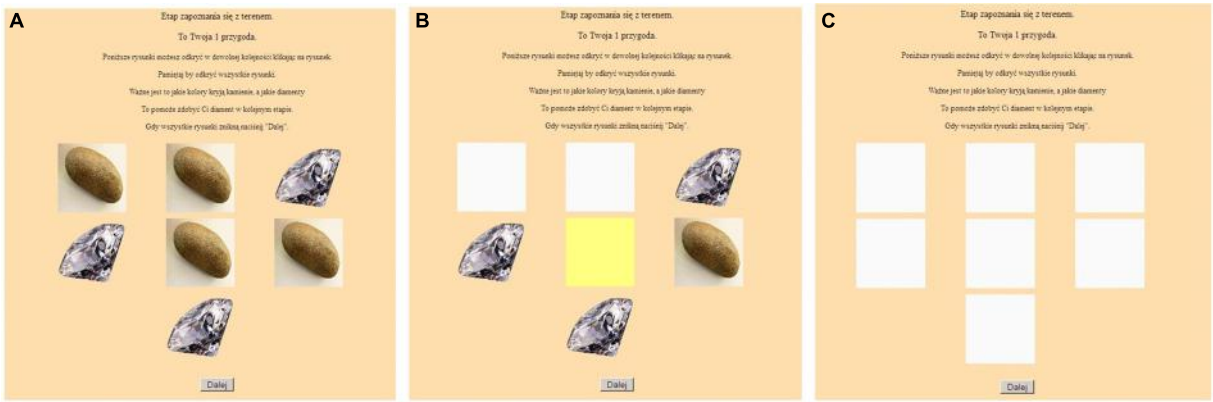

FIGURE 1 | The computer task: the learning stage – (A) before, (B) during, and (C) after turning over the cards.

will turn it over and reveal a color: green or yellow. Your task is to click on, i.e., turn over, all the cards to reveal colors on the back of the diamonds and stones. In the second stage, you should select the card with the color that has a diamond or a piece of amber underneath. You will take part in 16 such adventures." Subsequently, participants were asked if they understood the instructions. If so, they proceeded to perform the 16 tasks. Half of the adventures contained diamonds and stones, the other half, amber and glass, respectively. (For clarity, henceforth we only describe the method referring to diamonds and stones.)

Each adventure consisted of two stages.

The first stage was the learning stage, which was a simulation of natural sampling and was intended to develop intuition about Bayesian relationships. A participant was presented with seven cards showing valuable objects or worthless items on their faces (**Figure 1A**). The participant was instructed to turn over all of the cards in order to reveal the colors on their backs (**Figures 1B,C**). The person was to remember the colors associated with diamonds and stones, which would help them to acquire a diamond in the next step. Yellow and green colors were used for the back of the cards because these colors have relatively neutral emotional connotations (Karp and Karp, 1988).

In probability terms, a participant could learn:

- Prior occurrences of diamonds $[b/(b + c)]$ and stones $[c/(b + c)]$;
- Likelihood ratios for the backs of diamonds: green $[d/(d + e)]$, yellow $[e/(d + e)]$;
- Likelihood ratios for the backs of stones: green $[f/(f + g)]$, yellow $[g/(f + g)]$;
- Bayesian estimates of revealing a diamond for backs: green $[d/(d + f)]$, yellow $[e/(e + g)]$.

In the second stage of the adventure, participants chose between two differently colored cards (**Figure 2**). They received the following instructions: "Now you have a chance to find a diamond. There are two fields shown below, green, and yellow. One of them contains a desired diamond. Given what you have just learned, which color would you choose? Please select one card."

Choices satisfy Bayes' rule, when they are consistent with comparisons of the two Bayesian estimates shown above. We considered four strategies: Bayesian, pre-Bayesian, evidence-only and representativeness (Zhu and Gigerenzer, 2006). One binary choice would not allow us to discern between all the four strategies, as it has two alternatives only: two or more strategies could result in the same choice. Hence, strategies were inferred from pairs of adjacent adventures. To detect strategy use, we looked at eight pairs of adjacent adventures: 1 and 2, 3 and 4, 5 and 6, etc., up to 15 and 16. The second (even-numbered) task in a pair was determined by the first choice so as to allow distinct identification of the strategies used. For example, let us consider the first adventure, specified as $(d, e, f, g) = (2, 1, 3, 1)$. This means that we have the following cards: green-diamond (2), yellow-diamond (1), green-stone (3), yellow-stone (1). If a person used representative or evidence-only strategies they would select a green card. If they took a Bayesian or pre-Bayesian approach they would choose yellow. Suppose that a participant selected a yellow card in the first task. The second task in the pair was then specially matched to distinguish between use of a Bayesian or pre-Bayesian strategy. For example, it could take the form of a task specified as $(d, e, f, g) = (1, 2, 2, 2)$. If the participant chose a yellow card here, it was concluded that they used a Bayesian strategy. Similarly, other strategies were identified through matching the second task to the choice that was made in the first task in a pair. We used the following patterns of frequencies (d, e, f, g): P1 = (2,1,3,1), P2 = (1,2,1,3), P3 = (2,1,2,2), P4 = (2,1,1,3), P5 = (1,2,3,1), and P6 = (3,1,2,1). These served to construct the eight pairs of adventures as follows: pair I (P1 and: P3 when a yellow card was chosen in adventure P1 or P4 when a green card was chosen in P1), pair II (P1 and: P3 or P5), pair III (P1 and: P6 or P4), pair IV (P1 and: P6 or P5), pair V (P2 and: P3 or P4), pair VI (P2 and: P3 or P5), pair VII (P2 and: P6 or P4), and pair VIII (P2 and: P6 or P5). The program randomized the pairs and the on screen allocation of precious and invaluable items on different backgrounds. The content of adventures was also randomized and these consisted of one of the following two stimulus sets: (1)

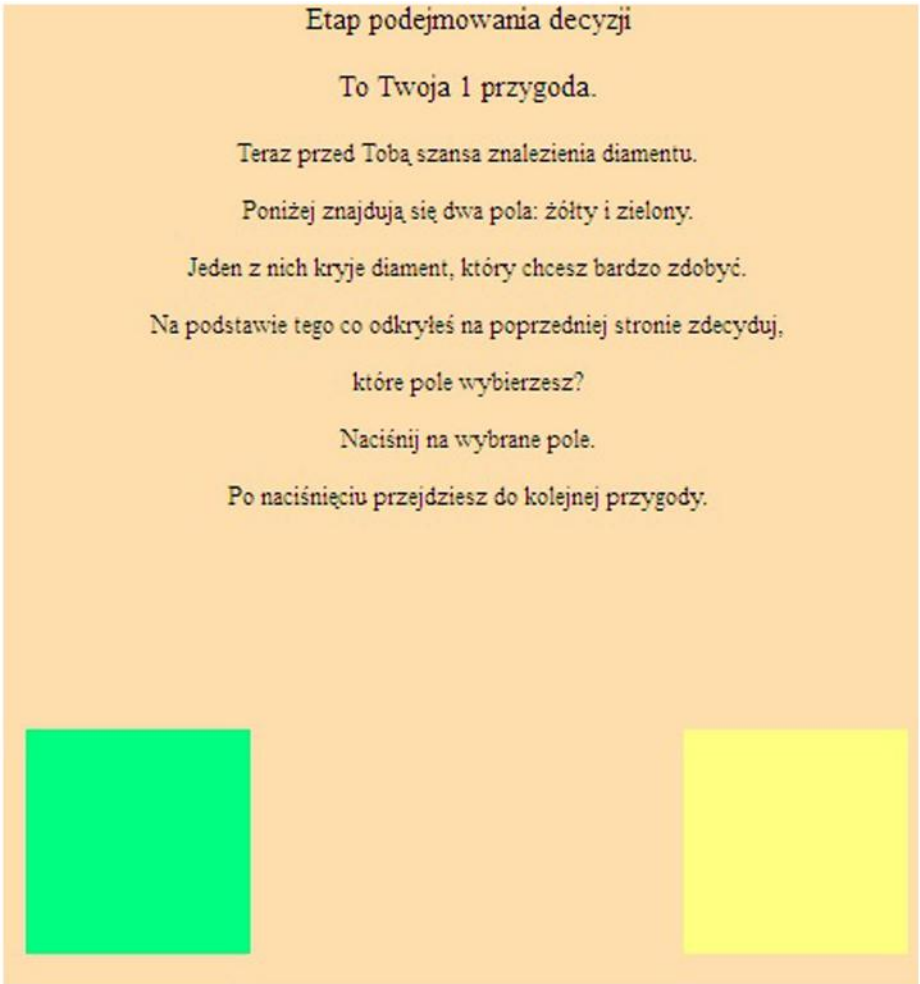

FIGURE 2 | The computer task: the color choice stage.

diamonds vs. stones in adventures 1–4 and 9–12, and ambers vs. pieces of glass in adventures 5-8 and 13–16, or (2) diamonds vs. stones in adventures 5–8 and 13–16, and ambers vs. pieces of glass in adventures 1–4 and 9–12.

Because we were looking for consistent application of the four strategies in eight pairs, scores for making choices conforming to Bayes' rule or other strategies ranged from 0 to 8, summing to 8. We assumed that participants used a strategy consistently and that any deviations from this strategy were accidental. However, it was possible that people might have applied different methods when solving different tasks (because of different task contents, practice, cognitive load, etc.). To provoke use of the same way of thinking in all of the tasks, we provided no feedback during testing so that participants would not learn from practice. Thus, we did not suggest to participants which data they should take into consideration. All tasks were homogeneous in terms of content, format, and difficulty. To minimize cognitive load we limited the learning phase to clicking on seven pictures only and required every adventure to be solved separately. We asked participants to complete all of the tasks at once to prevent any change in skills. We attributed all inconsistencies in responses and strategies applied to random noise and errors.

The Studies 1 and 2 experiments were approved by Scientific Research Ethics Committee at the Faculty of Psychology, University of Warsaw, and informed consent was obtained from all subjects.

Participants and Procedure

A stratified sample of $N = 60$ students aged 20–35 ($M = 24.58$ years, $SD = 3.16$) volunteered for the study. Participants were equally distributed with regard to gender and type of education (humanities and pure sciences). Individual interviews took place at the University of Warsaw and Warsaw University of Technology. The study was presented as a computer game involving gathering precious items. Completing all of the tasks took about 15 min. Participation was anonymous and not rewarded. At the end, participants were informed about their scores. We then acquainted participants with the actual objectives of the study.

Results and Interpretation

Table 1 shows how often each strategy was applied.

Choices conforming to Bayes' rule were more common than they would be at random [$M = 4.58$, SD $= 2.42$, test value: $\mu = 2$, $t(59) = 8.27$, $p < 0.001$, $d = 1.067$; Scaled JZS Bayes Factor $B = 5.15 \times 10^8$, supporting $\mu > 2$]. Therefore, the weak version of the hypothesis (Hb) was supported. The Bayesian strategy was dominant and was used in slightly more than half of the cases, however, test statistics were non-significant [test value: $\mu = 4$, $t(59) = 1.87$, $p = 0.067$, $d = 0.241$, with a non-decisive Scaled JZS $B = 1.283$]. Thus, the strong version of the hypothesis (Ha) was not supported.

Participants' choices conformed to Bayes' rule in a majority of cases (57%, $M = 4.58$ out of 8), showing that the strategy was used more often than by chance. Furthermore, it was more popular than all the other strategies taken together. The weak hypothesis was supported, but the results involving the strong hypothesis were marginally non-significant. However, the natural sampling procedure demanded that participants computed and compared natural frequencies. This makes natural sampling tasks involving choices potentially more intellectually demanding than pure natural frequency problems. One would therefore expect a greater percentage of fallacious answers when natural sampling is used.

While the adopted methodology resembled natural sampling, it obscured the process of inference underlying choices. A decision based on experience has four phases: (1) gathering information (counting objects); (2) building a mental representation (such as classes of objects and their proportions); (3) processing of information using a choice mechanism (comparison of estimates); (4) making a final selection. Only information gathering and the final decision are external, observable events (Camilleri and Newell, 2009). Therefore, as our results might have appeared to be rather optimistic, we decided to replicate Study 1 but asking participants how they solved the problems in more detail.

TABLE 1 | Strategies applied for Bayesian problems in Study 1.

Strategies	Descriptive Statistics ($N = 60$)			
	Min	*Max*	*M*	SD
Bayesian	0	8	4.58	2.42
Pre-Bayesian	0	4	0.82	1.13
Representativeness	0	7	2.15	2.07
Evidence-only	0	5	0.45	1.06

Study 2: Replication of Study 1 with Verbal Protocols

The goal of Study 2 was to replicate Study 1 so as to identify strategies applied in Bayesian tasks more directly. We utilized a process tracing method (Baron, 1994, pp. 19–24). The classical process tracing approach specifies that participants should not be requested to justify their decisions (Nisbett and Wilson, 1977; Ericsson and Simon, 1980). However, participants should easily explain their choices, since the contents of tasks included simple notions, numbers, and computations.

Materials and Methods

Study 2 was intended to generate results comparable to those from Study 1. The study used the same set of computer tasks as Study 1. After completing the tasks, participants were asked to solve an additional Bayesian exercise. This exercise reproduced a computer task, but was conducted using paper cards. The experimenter presented seven cards with diamonds and stones on and then asked a participant to turn over the cards. After they were all turned over, the cards were taken away and two cards were presented: one yellow and one green. Before uncovering one of them, the participant was asked about the method they used to solve the exercise. The experimenter refrained from providing any suggestions or clues as to how to perform the task or make any computations. Thus, the method applied here differed from the "write aloud" protocols used by Gigerenzer and Hoffrage (1995). At the end of the procedure, the experimenter classified the participant's answer using the coding list presented in **Table 2**. For example, where a participant compared the natural frequencies of differently colored cards to their total number the experimenter registered this as an evidence-only strategy.

Participants and Procedure

A sample of $N = 76$ students aged 18–31 ($M = 23.82$ years, SD $= 2.17$) volunteered for the study. Participants were equally distributed into four cells ($n = 19$ each) with regard to gender

TABLE 2 | Coding strategies identified in verbal protocols on the paper task.

Verbal explanation	Interpreted as using the strategy
Comparing relative or absolute frequencies of yellow and green diamonds: d/b vs. e/b or d vs. e	Representativeness
Comparing relative or absolute frequencies of yellow and green cards: $d + f$ vs. $e + g$ or $(d + f)/a$ vs. $(e + g)/a$	Evidence-only
Comparing the relationship of the number of cards with diamonds to the number of cards with defined colors: $(d + e)/(d + f)$ vs. $(d + e)/(e + g)$	Pre-Bayesian
Comparing empirical probabilities of cards with diamonds among yellow cards with empirical probabilities of cards with diamonds among green cards: $d/(d + f)$ vs. $e/(e + g)$	Bayesian
Comparing numbers of cards with diamonds and stones: b vs. c	Conservatism
Other explanations (mixed strategies, guessing, intuition, etc.)	Mixed/guessing/other

and type of education. We applied the same procedure as in Study 1 but added the paper task. The experimenter presented the computer-based tasks from Study 1, followed by the additional exercise, individually to each participant.

Results

Bayesian and Other Strategies

The Bayesian strategy was applied significantly more often than would occur randomly [test value: $\mu = 2$, $t(75) = 7.41$, $p < 0.001$, $d = 0.850$; Scaled JZS $B = 6.33 \times 10^7$, supporting $\mu > 2$ – see **Table 3**]. This strategy again dominated, being utilized in more than half of the cases. Nevertheless, the extent to which use of the strategy exceeded half of the cases was non-significant [test value: $\mu = 4$, $t(75) = 0.745$, $p > 0.10$, $d = 0.850$, Scaled JZS $B = 1.465$ was not decisive]. Thus, again there was support for the weak criterion (Hb), but the strong criterion (Ha) went unsupported. Hence, Study 2 replicated the results of Study 1.

Verbal Protocol vs. Computer-Based Tasks

Participants' verbal explanations revealed a new, quite frequently used strategy (32% participants in the whole sample: 18 out of 33 who used the Bayesian strategy in computer tasks, and 8 out of 11 who used heuristics).

The new strategy is different from the strategies listed in **Table 2**. This new strategy included comparing the number of yellow (green) cards among diamonds with the yellow (green) cards among stones. Using the notation we adopted, for yellow this would be: $d/(d + e)$ vs. $f/(f + g)$, and for green: $e/(d + e)$ vs. $g/(f + g)$. Using this strategy does not require inverse thinking about conditions and computing P(H|D) when P(D|H) is given. Intriguingly, this new strategy produces choices that are always the same as choices based on using Bayes' rule. Comparing $d/(d + e)$ with $f/(f + g)$ is equivalent mathematically with comparing $d \times (f + g)$ with $f \times (d + e)$, and subsequently: $(d \times g + d \times f)$ with $(e \times f + d \times f)$; $d \times g$ with $e \times f$; $(d \times g + d \times e)$ with $(e \times f + d \times e)$; $d \times (e + g)$ with $e \times (d + f)$, and finally $d/(d + f)$ with $e/(e + g)$. This last comparison represents the Bayesian strategy.

Most participants (57 out of 76, i.e., 75%) used consistently algorithmic (Bayesian or the new strategy) or fallacious strategies in both the computer and paper card tasks (**Table 4**).

Thirty-three out of 41 participants (80%), whose dominant strategy was the Bayesian strategy in computer tasks, used the Bayesian strategy or the non-inverse strategy in the paper tasks. Twenty-four out of 35 (69%) used other strategies in both types of tasks. Consistency in using dominant strategies in the computer-based tasks and analogous strategies in paper excercises was moderate [$\chi^2(1, N = 76) = 18.64$, $p < 0.001$, $\varphi = 0.495$]. Summing up, Study 2 confirmed the results of Study 1, showing that most choices were consistent with Bayes' rule. However, they were the result of using of not only Bayes' strategy, but also the new, non-inverse strategy.

TABLE 3 | Strategies applied for Bayesian problems in Study 2.

Strategies	Descriptive statistics ($N = 76$)			
	Min	*Max*	*M*	SD
Bayesian	0	8	4.22	2.62
Pre-Bayesian	0	4	0.96	1.08
Representativeness	0	8	2.34	2.24
Evidence-only	0	4	0.47	0.92

TABLE 4 | Dominant strategies in computer tasks vs. strategies used in the paper task in Study 2.

Dominant strategies in computer tasks	Verbal reports in the paper tasks		Total
	Bayesian or the new strategy	Other strategies	
Bayesian strategy	33 (80%)	8 (20%)	41 (100%)
Other strategies	11 (31%)	24 (69%)	35 (100%)
Total	44 (58%)	32 (42%)	76 (100%)

Study 3 (An Analytical Study)

The Bayesian strategy and the new non-inverse strategy identified in Study 2 provide answers that are always correct in terms of Bayes' rule. However, people may compromise between the effort and time needed to make consistently correct choices and the practical convenience of making fast and frugal choices. In this section, we investigate how often using fallacious strategies (representativeness, evidence-only and pre-Bayesian strategies) leads to the same choices as does using Bayes' rule. We analyze strategies with regard to (1) different frequencies expressing decision-makers' natural sampling experiences and (2) different base rates, arbitrarily defined as rare [$P(H) \leq 0.25$], frequent [$P(H) \geq 0.75$], and medium [$0.25 < P(H) < 0.75$].

Method

Let us start with an example. Consider an elementary situation $(d, e, f, g) = (4, 1, 1, 1)$, where d denotes number of cards with a diamond on its face and a green back, e – diamond-yellow, f – stone-green, and g – stone-yellow, respectively. Using the Bayesian strategy, a person should choose a green card to reveal a diamond, because: $d/(d + f) = 4/(4 + 1) > e/(e + g) = 1/(1 + 1)$. The same answer would result from using the representativeness strategy [$d/b = 4/5 > e/b = 1/5$], or the evidence-only strategy: $(d + f)/a = (4 + 1)/7 > (e + g)/a = (1 + 1)/7$. The pre-Bayesian strategy would render solutions greater than one for yellow cards, $(d + e)/(e + g) = (4 + 1)/(1 + 1) = 5/2$. In such cases, when the probability estimates exceed one, we consider the strategy inapplicable.

We wanted to understand how often non-Bayesian strategies return results as good as the correct, Bayesian strategy. We generated all combinations of (d, e, f, g) for sampling volumes $d + e + f + g = a$ ranging from 5 to 50, for $d, e, f, g > 0$ (every combination of data and hypotheses was experienced at least once). For example, **Table 5** shows prescriptions for a choice in all twenty possible elementary situations when $a = 7$. Here, D_1 means reversing a green card and D_2 means

TABLE 5 | Conformity of the heuristic strategies to Bayes' strategy in choice prescription.

a	*d*	*e*	*f*	*g*	Bayesian	Representativeness		Evidence-only		Pre-Bayesian	
					Choice	Choice	Conformity	Choice	Conformity	Choice	Conformity
7	1	1	1	4	D_1	Any	No	D_2	No	D_1	Yes
7	1	1	2	3	D_1	Any	No	D_2	No	D_1	Yes
7	1	1	3	2	D_2	Any	No	D_1	No	D_2	Yes
7	1	1	4	1	D_2	Any	No	D_1	No	D_2	Yes
7	1	2	1	3	D_1	D_2	No	D_2	No	n/a	
7	1	2	2	2	D_2	D_2	Yes	D_2	Yes	D_1	No
7	1	2	3	1	D_2	D_2	Yes	D_1	No	D_2	Yes
7	1	3	1	2	D_2	D_2	Yes	D_2	Yes	n/a	
7	1	3	2	1	D_2	D_2	Yes	D_2	Yes	n/a	
7	1	4	1	1	D_2	D_2	Yes	D_2	Yes	n/a	
7	2	1	1	3	D_1	D_1	Yes	D_2	No	D_1	Yes
7	2	1	2	2	D_1	D_1	Yes	D_1	Yes	D_2	No
7	2	1	3	1	D_2	D_1	No	D_1	No	n/a	
7	2	2	1	2	D_1	Any	No	D_2	no	n/a	
7	2	2	2	1	D_2	Any	No	D_1	no	n/a	
7	2	3	1	1	D_2	D_2	Yes	D_2	Yes	n/a	
7	3	1	1	2	D_1	D_1	Yes	D_1	Yes	n/a	
7	3	1	2	1	D_1	D_1	Yes	D_1	Yes	n/a	
7	3	2	1	1	D_1	D_1	Yes	D_1	Yes	n/a	
7	4	1	1	1	D_1	D_1	Yes	D_1	Yes	n/a	
Conformity:						12/20 = 60%		10/20 = 50%		6/8 = 75%	

reversing a yellow card. It turned out that if a was 7, then: (1) the representativeness strategy conforms to Bayes' rule in 60% of situations; (2) evidence-only – in 50%; (3) pre-Bayesian – in 75% (out of situations where the strategy is applicable).

Results and Interpretation

The analysis showed that the higher the volume of sampling *a*, the more stable is the percentage of elementary situations in which using a given strategy leads to choices conforming to Bayes rule (see **Figure 3**). The average number of Bayesian solutions returned by a strategy is: (a) representativeness – 73%, (b) evidence-only – 50%, (c) pre-Bayesian – 63%.

The representativeness strategy is effective for high base rates and small natural sampling sizes (**Figure 4**). Specifically, when $a \leq 11$ and the base rate is $b/a = (d + e)/(d + e + f + g) \geq 0.75$, the representativeness strategy always produces choices conforming to Bayes' rule. If the base rate exceeds 0.75, the representativeness strategy returns correct choices in no less than 77.9% of cases. However, if the base rate is low ($b/a \leq 0.25$), even if the size is high ($a > 11$), choices conforming to Bayes' rule are generated at a rate between 42.9% and 67.6%. In contrast, at a low volume of sampling ($a \leq 11$) and low base rate ($b/a \leq 0.25$) it produces optimal selections in only 20% or fewer situations.

The evidence-only strategy returns choices conforming to Bayes' rule in 50% of cases at moderate base rates (**Figure 5**). If the base rate (b/a) exceeds 0.75, the strategy produces correct answers in 72.6% or more of cases. However, when the base rate is lower than 0.25, it produces choices conforming to Bayes' rule with a probability of 26.5% or less. We also noticed that if $a \leq 11$ and $b/a \geq 0.75$ the evidence-only strategy is always right. Conversely, for $b/a \leq 0.25$ it renders correct answers in 20% or fewer situations.

By definition, the pre-Bayesian strategy always gives opposite answers to the evidence-only strategy (**Figure 6**) and, indeed, we observed its diametrically opposite behavior for all size – base rate combinations. A decision maker should understand that probabilities do not exceed one, i.e., $(d + e)/(d + f) \leq 1$ and $(d + e)/(e + g) \leq 1$. This implies $2(d + e) \leq (d + f + e + g)$, $2b \leq a$ and $b/a \leq 0.5$, and means that the strategy is not applicable for base rates exceeding 1/2. With these assumptions, the strategy renders choices conforming to Bayes' rule with a probability of 56.0% for medium base rates, and 72.6% for low base rates.

Summing up, the representativeness and evidence-only strategies return choices conforming to Bayes' rule with very high probabilities if base rates are high and the natural sampling size is low. The pre-Bayesian strategy turned out to be far less efficient.

Discussion

The first goal of our studies was to find out how often choices in elementary situations satisfy Bayes' rule, if probabilistic information is acquired through natural sampling. Many studies on Bayesian reasoning have expected that solitary probability estimation should follow the rule. We extended this expectation

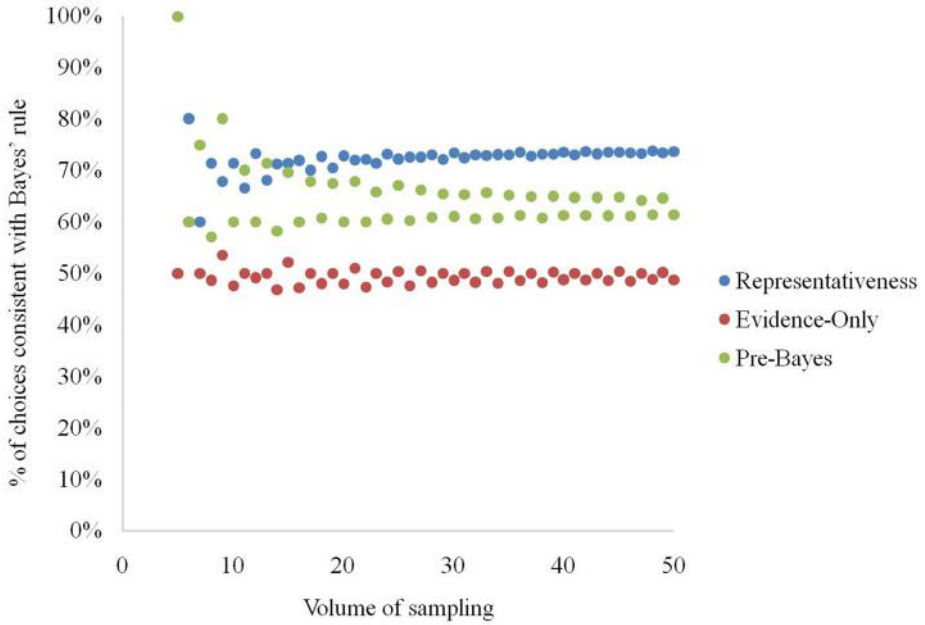

FIGURE 3 | Natural sampling volume and percentage of elementary situations in which the strategies conform to Bayes' rule in producing choices.

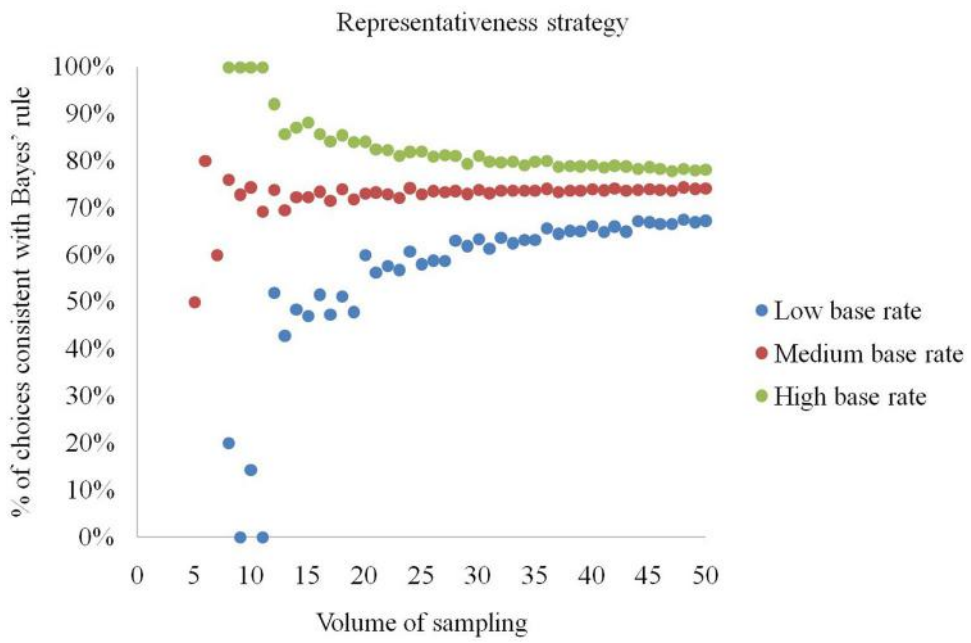

FIGURE 4 | Percentage of elementary situations in which the representativeness strategy produces choices consistent with Bayes' rule at low, medium and high base rates.

to choices, however, we did not require participants to evaluate chances, we only asked them make choices.

Our studies confirmed that most choices satisfied Bayes' rule. Overall, the results were consistent with studies in which the application of natural frequency formats has improved the proportion of Bayesian responses, varying in the range from 31 to 72% (as compared by Barbey and Sloman, 2007), or as high as 77% (in the group of adults investigated by Zhu and Gigerenzer, 2006). One could then conclude that natural sampling facilitates Bayesian inference in elementary situations. Participants were allowed to uncover cards at their own pace and using their own sequences. They discovered connections between objects and

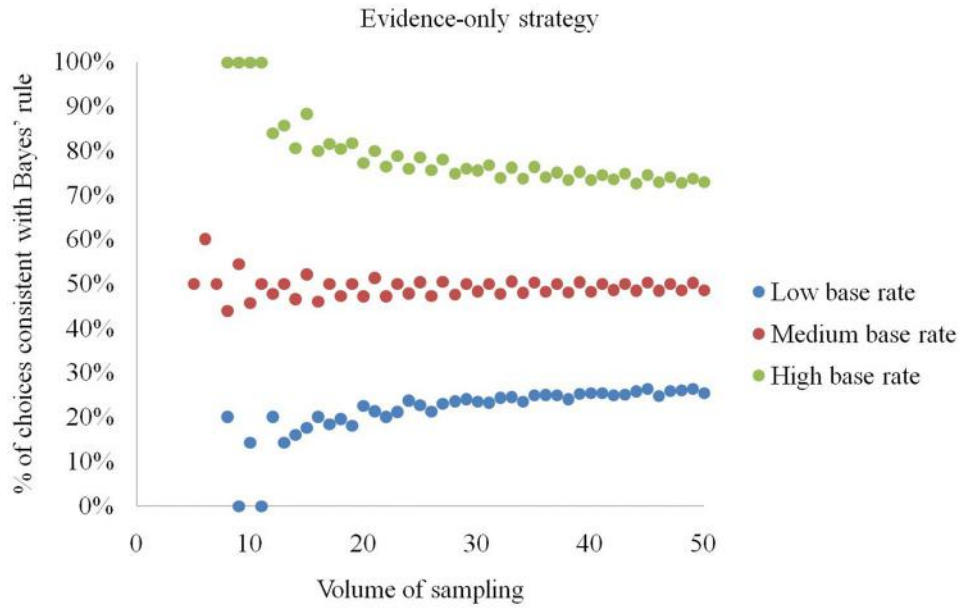

FIGURE 5 | Percentage of elementary situations in which the evidence-only strategy produces choices consistent with Bayes' rule at low, medium and high base rates.

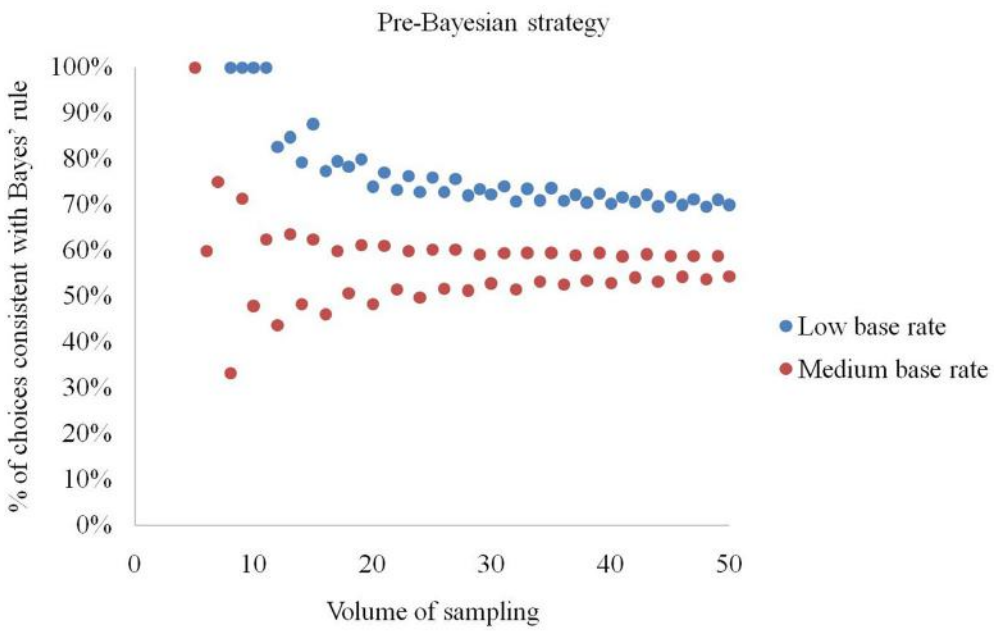

FIGURE 6 | Percentage of elementary situations in which the pre-Bayesian strategy produces choices consistent with Bayes' rule at low and medium base rates.

colors on their own terms. As we gave no suggestions about how to solve the problems, participants could utilize their own estimates or impressions. Moreover, participants operated on cards at both stages of the task. This compatibility between presented data and answer format could also have enhanced performance (as concluded by Ayal and Beyth-Marom, 2014). Because these results seemed rather optimistic with regard to tasks' complexity, so we decided to replicate the study adding verbal protocols, which revealed the strategies used more directly.

Although Study 2 replicated the results of Study 1, it turned out that a considerable number of correct choices resulted not from using Bayes' rule but from a new non-inverse strategy. This

method always renders the same answers as the Bayesian strategy in elementary situations and was therefore indistinguishable if only choices were examined. The non-inverse strategy involves computing likelihood ratios, P(D|H) and P(D|not-H), instead of Bayesian posterior probabilities, $P(H|D_1)$ and $P(H|D_2)$. In other words, the strategy focuses on a given datum (e.g., the green back of a card) and determines whether it is more characteristic for the hypothesis, H (e.g., a diamond), or for the alternative hypothesis, not-H (a stone). Usually, sticking to likelihood ratios or confusing them with posterior probabilities in Bayesian problems is considered fallacious and is called "an inverse fallacy" (Villejoubert and Mandel, 2002; Mandel, 2014). The confusion of conditions is indeed erroneous, e.g., believing that if most amber is found on yellow beaches then you can find amber on a majority of yellow beaches. However, replacement of both $P(H|D_1)$ and $P(H|D_2)$, with both $P(D_1|H)$ and $P(D_1|\text{not-H})$, or both $P(D_2|H)$ and $P(D_2|\text{not-H})$ is not fallacious. Here, resulting choices are always consistent with Bayes' rule. The non-inverse strategy is mathematically equivalent to the calculation of the difference P(D|H)–P(D|not-H). This computation was observed in studies by Gigerenzer and Hoffrage (1995), who named it a likelihood subtraction method. These authors concluded that users of this strategy neglect base rate information. However, this might be true only when likelihood ratios are input data, as is the case in typical Bayesian tasks. When natural sampling is applied, as in our studies, people must consider base frequencies for estimating likelihood ratios on their own. This finding supports the proposition that learning from direct experience reduces base-rate neglect (Koehler, 1996; Hertwig and Ortmann, 2001).

Study 3 showed that non-Bayesian, heuristic strategies handled tasks quite well in elementary situations under certain specific circumstances. At low base rates, the pre-Bayesian strategy suggested choices that satisfy Bayes' rule in most cases at a low volume of natural sampling. The representativeness and evidence-only strategies turned out to be successful under the specific conditions of high base-rates of the distinct hypothesis and low natural sampling sizes (few cards). These findings may explain some difficulties and fallacious propensities in solving Bayesian tasks described in the literature. What would happen, for instance, in the taxi cab problem if, instead of asking participants to give a probability that the taxi cab was blue, we asked them for the probability that the taxi cab was green, given that the witness claimed this to be the case? The findings of such a study would not be very impressive. Fallacious strategies would provide the same interpretation as Bayes' rule, which would give a 95.8% probability. A conservative strategy would return an estimate of 85%, representativeness – 80%, evidence-only – 71%, and pre-Bayesian – 83.5%. Any strategy would indicate that it was most probably a green cab if a witness claimed it to be so. Thus, it is not necessary to use Bayes' rule to make a correct decision or judgment based on probability magnitude.

We would like to emphasize that our findings are limited to elementary situations only. Such a limited, local application of strategies and heuristics is consistent with an ecological view. Gigerenzer (1991) pointed out that it is crucial to take into account the environment when one wants to evaluate the approach applied. It is also in line with probabilistic functionalism, which suggests that not using bookish methods for their own sake, but using any methods for achieving goals in the environment, drives human behavior (Pleskac and Hertwig, 2014). The tasks required the selection of green or yellow cards in order to maximize the probability of receiving a diamond instead of a stone.

A natural extension of our studies would be to investigate larger natural sampling sizes and exercises involving more data and more hypotheses. In such a situation the non-inverse strategy does not generalize and would be misleading. Also, heuristic strategies would be likely to be far less efficient in such complex, non-elementary situations.

We are quite pessimistic about humans' ability to solve such complex problems in a Bayesian way. First, people reveal little interest in gathering complete information on probabilities in naturalistic risky tasks (Huber et al., 1997; Tyszka and Zaleśkiewicz, 2006). Second, if the sample size were increased, working memory boundaries would be exceeded (Anderson, 2000). Longer sampling sequences would probably increase computational complexity, decrease participants' performance, and provoke them to make more use of various heuristics. The assumption that people use a given strategy consistently within a set of tasks (or at least within pairs of tasks) is challenging and difficult to maintain. This assumption was the main limitation of our studies, but it was necessary to infer strategies from choices indirectly. We tried to minimize the risk of participants using various strategies by presenting only seven cards in a task with homogeneous contents, and giving no feedback. On the one hand, if the assumption is rejected, the problem remains as to how to reveal thinking underlying choices directly, and – at the same time – not to tell participants which chances should be evaluated and how. On the other hand, the assumption is problematic because factors such as skills, cognitive load, learning effects, more differentiated contents, etc. would likely entail applying different heuristics, particularly in more complex tasks.

In analyzing choices in elementary situations we adopted a narrow definition of Bayesian inference as choices or probability evaluations conforming to Bayes' rule (similarly to other psychological studies investigating Bayesian reasoning). However, Bayesian inference might be understood as the general process of using new information to revise evaluations of likelihoods of events with known prior base rates (Brase and Hill, 2015). In particular, this describes Bayesian analysis of decision problems incorporated in subjective expected utility theory (SEUT, Savage, 1954; Giocoli, 2013; Karni, 2013). According to this perspective, a Bayesian decision-maker's subjective beliefs are expressed with probabilities which are updated in line with Bayes rule as new information is gathered. Hypothesized outcomes (e.g., diamonds and stones in our studies) are characterized by their utilities [e.g., $U(H_1)$, $U(H_2)$, $U(H_1) > U(H_2)$]. The decision maker maximizes the subjective expected utility (SEU) of choice options, combining the subjective probabilities and utilities of outcomes. If the choices are made in elementary situations, as in our studies, maximizing SEU reduces to choosing the option characterized by the higher posterior chance [$SEU(D_1) > SEU(D_2)$ when $P(H_1|D_1) \times U(H_1)+P(H_2|D_1) \times U(H_2) > P(H_1|D_2) \times U(H_1) + P(H_2|D_2) \times U(H_2)$], and

$[P(H_1|D_1) - P(H_1|D_2)] \times [U(H_1) - U(H_2)] > 0$, and subsequently $[P(H_1|D_1) - P(H_1|D_2) > 0]$. However, extending the analysis of choice to more complex situations with more than two hypothesized outcomes (e.g., diamonds, stones, and graphite) entails incorporation of their utilities into the analysis. Here, choice does not reduce to comparing probabilities, and differences among utilities influence the final choice, which is made by maximizing SEU.

Summing up, people performed well in the Bayesian exercises involving natural sampling in elementary situations in our studies. However, correct Bayesian choices can result from using non-Bayesian methods, such as the non-inverse strategy identified in our studies. What is more, even fallacious heuristics produce satisficing choices reasonably often under specific circumstances. Hence, Bayesian inference turns out to be unnecessary in making choices satisfying Bayes' rule in elementary situations.

Acknowledgments

We would like to thank Prof. Tadeusz Tyszka, Dr. Łukasz Markiewicz, and Dr. Janina Pietrzak for their helpful comments. This work was supported by grants BST 1250 34 and BST 1712 26 given to AD at the University of Warsaw.

References

Anderson, J. R. (2000). *Learning and Memory: An Integrated Approach,* 2nd Edn. New York: Wiley.

Ayal, S., and Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgm. Decis. Mak.* 9, 226–242.

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–254. doi: 10.1017/S0140525X07001653

Baron, J. (1994). *Thinking and Deciding,* 2nd Edn. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511840265

Brase, G. L. (2008). Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol.* 23, 369–381. doi: 10.1002/acp.1460

Brase, G. L., Cosmides, L., and Tooby, J. (1998). Individuation, counting, and statistical inference: the role of frequency and whole-object representations in judgment under uncertainty. *J. Exp. Psychol. Gen.* 127, 3–21. doi: 10.1037/0096-3445.127.1.3

Brase, G. L., and Hill, W. T. (2015). Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Front. Psychol.* 6:340. doi: 10.3389/fpsyg.2015.00340

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychol. Rev.* 50, 255–272. doi: 10.1037/h0060889

Camilleri, A. R., and Newell, B. R. (2009). The role of representation in experience-based choice. *Judgm. Decis. Mak.* 4, 518–529.

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

Dhami, M. K., Hertwig, R., and Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychol. Bull.* 130, 959–988. doi: 10.1037/0033-2909.130.6.959

Eddy, D. (1982). "Probabilistic reasoning in clinical medicine: problems and opportunities," in *Judgment Under Uncertainty: Heuristics and Biases,* eds D. Kahneman, P. Slovic, and A. Tversky (Cambridge: Cambridge University Press), 249–267. doi: 10.1017/CBO9780511809477.019

Edwards W. (1968). "Conservatism in human information processing," in *Formal Representation of Human Judgment*, ed. B. Kleinmuntz (New York: Wiley), 17–52.

Ericsson, K. A., and Simon, H. A. (1980). Verbal reports as data. *Psychol. Rev.* 87, 215–251. doi: 10.1037/0033-295X.87.3.215

Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond heuristics and biases. *Eur. Rev. Soc. Psychol.* 2, 83–115. doi: 10.1080/14792779143000033

Gigerenzer, G. (1998). "Ecological intelligence. An adaptation for frequencies," in *The Evolution of Mind,* eds D. D. Cummins and C. Allen (New York: Oxford University Press), 9–29.

Gigerenzer, G. (2004). "Fast and frugal heuristics: the tools of bounded rationality," in *Blackwell Handbook of Judgment and Decision Making,* eds D. J. Koehler and N. Harvey (Malden, MA: Blackwell Publishing), 62–88. doi: 10.1002/9780470752937.ch4

Gigerenzer, G. (2008). Why heuristics work. *Perspect. Psychol. Sci.* 3, 20–29. doi: 10.1111/j.1745-6916.2008.00058.x

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., Hoffrage, U., and Ebert, A. (1998). AIDS counseling for low-risk clients. *AIDS Care* 10, 197–211. doi: 10.1080/09540129850124451

Gigerenzer, G., Todd, P. M., and the ABC Research Group. (1999). *Simple Heuristics that Make Us Smart*. New York: Oxford University Press.

Giocoli, N. (2013). From wald to savage: homo economicus becomes a Bayesian statistician. *J. Hist. Behav. Sci.* 49, 63–95. doi: 10.1002/jhbs.21579

Girotto, V., and Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition* 78, 247–276. doi: 10.1016/S0010-0277(00)00133-5

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychol. Sci.* 15, 534–539. doi: 10.1111/j.0956-7976.2004.00715.x

Hertwig, R., and Erev, I. (2009). The description–experience gap in risky choice. *Trends Cogn. Sci.* 13, 517–523. doi: 10.1016/j.tics.2009.09.004

Hertwig, R., and Ortmann, A. (2001). Experimental practices in economics: a methodological challenge for psychologists? *Behav. Brain Sci.* 24, 383–403. doi: 10.1037/e683322011-032

Hoffrage, U., Hafenbrädl, S., and Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Front. Psychol.* 6:642. doi: 10.3389/fpsyg.2015.00642

Huber, O., Wider, R., and Huber, O. W. (1997). Active information search and complete information presentation in naturalistic risky decision tasks. *Acta Psychol.* 95, 15–29. doi: 10.1016/S0001-6918(96)00028-5

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1007/978-94-010-2288-0-3

Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychol. Rev.* 80, 237–251. doi: 10.1037/h0034747

Karni, E. (2013). Bayesian decision theory with action-dependent probabilities and risk attitudes. *Econ. Theory* 53, 335–356. doi: 10.1007/s00199-012-0692-4

Karp, E. M., and Karp, H. B. (1988). Color associations of male and female fourth-grade school children. *J. Psychol.* 122, 383–388. doi: 10.1080/00223980.1988.9915525

Kleiter, G. D. (1994). "Natural sampling: rationality without base rates," in *Contributions to Mathematical Psychology, Psychometrics, and Methodology,* eds G. H. Fischer and D. Laming (New York: Springer), 375–388. doi: 10.1007/978-1-4612-4308-3-27

Koehler, J. J. (1996). The base rate fallacy reconsidered: descriptive, normative, and methodological challenges. *Behav. Brain Sci.* 19, 1–17. doi: 10.1017/S0140525X00041157

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Front. Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Front. Psychol.* 6:387. doi: 10.3389/fpsyg.2015.00387

Mellers, B., and McGraw, A. P. (1999). How to improve Bayesian reasoning: comment on Gigerenzer and Hoffrage (1995). *Psychol. Rev.* 106, 417–424. doi: 10.1037/0033-295x.106.2.417

Nance, D. A., and Morris, S. B. (2005). Juror understanding of DNA evidence: an empirical assessment of presentation formats for trace evidence with a relatively small random-match probability. *J. Legal Stud.* 34, 395–444. doi: 10.1086/428020

Navarrete, G., Correia, R., and Froimovitch, D. (2014). Communicating risk in prenatal screening: the consequences of Bayesian misapprehension. *Front. Psychol.* 5:1272. doi: 10.3389/fpsyg.2014.01272

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychol. Rev.* 84, 231–259. doi: 10.1037/0033-295X.84.3.231

Over, D. E. (2004). "Rationality and the normative/descriptive distinction," in *Blackwell Handbook of Judgment and Decision Making*, eds D. J. Koehler and N. Harvey (Malden, MA: Blackwell Publishing), 3–18. doi: 10.1002/9780470752937.ch1

Phillips, L. D., and Edwards, W. (1966). Conservatism in a simple probability inference task. *J. Exp. Psychol.* 72, 346–354. doi: 10.1037/h0023653

Pleskac, T. J., and Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *J. Exp. Psychol.* 143, 2000–2019. doi: 10.1037/xge0000013

Rakow, T., and Newell, B. R. (2010). Degrees of uncertainty: an overview and framework for future research on experience-based choice. *J. Behav. Decis. Mak.* 14, 1–14. doi: 10.1002/bdm.681

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Wiley.

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol.* 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Simon, H. A. (1955). A behavioral model of rational choice. *Q. J. Econ.* 69, 99–118. doi: 10.2307/1884852

Simon, H. A. (1956). Rational choice and the structure of environment. *Psychol. Rev.* 63, 129–138. doi: 10.1037/h0042769

Sirota, M., Juanchich, M., and Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychon. Bull. Rev.* 21, 198–204. doi: 10.3758/s13423-013-0464-6

Sloman, S. A., Over, D., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Process.* 91, 296–309. doi: 10.1016/S0749-5978(03)00021-9

Tversky, A., and Kahneman, D. (1982). "Evidential impact of base rates," in *Judgment Under Uncertainty: Heuristics and Biases,* eds D. Kahneman, P. Slovic, and A. Tversky (Cambridge: Cambridge University Press), 153–160. doi: 10.1017/CBO9780511809477.011

Tyszka, T., and Zaleśkiewicz, T. (2006). When does information about probability count in choices under risk? *Risk Anal.* 26, 1623–1636. doi: 10.1111/j.1539-6924.2006.00847.x

Villejoubert, G., and Mandel, D. R. (2002). The inverse fallacy: an account of deviations from Bayes theorem and the additivity principle. *Mem. Cogn.* 30, 171–178. doi: 10.3758/BF03195278

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: frequency or nested sets? *Exp. Psychol.* 50, 97–106. doi: 10.1026//1618-3169.50.2.97

Zhu, L., and Gigerenzer, G. (2006). Children can solve Bayesian problems: the role of representation in mental computation. *Cognition* 98, 287–308. doi: 10.1016/j.cognition.2004.12.003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 April 2015
doi: 10.3389/fpsyg.2015.00398

Uncertain deduction and conditional reasoning

Jonathan St. B. T. Evans[1]*, *Valerie A. Thompson*[2] and *David E. Over*[3]*

[1] School of Psychology, University of Plymouth, Plymouth, UK, [2] Department of Psychology, University of Saskatchewan, Saskatoon, SK, Canada, [3] Department of Psychology, Durham University, Durham, UK

There has been a paradigm shift in the psychology of deductive reasoning. Many researchers no longer think it is appropriate to ask people to assume premises and decide what necessarily follows, with the results evaluated by binary extensional logic. Most every day and scientific inference is made from more or less confidently held beliefs and not assumptions, and the relevant normative standard is Bayesian probability theory. We argue that the study of "uncertain deduction" should directly ask people to assign probabilities to both premises and conclusions, and report an experiment using this method. We assess this reasoning by two Bayesian metrics: probabilistic validity and coherence according to probability theory. On both measures, participants perform above chance in conditional reasoning, but they do much better when statements are grouped as inferences, rather than evaluated in separate tasks.

Keywords: uncertain premises, conditional reasoning, new paradigm psychology of reasoning, p-validity, coherence, explicit inference, fallacy

Edited by:
David R. Mandel, Defence Research and Development Canada Toronto, Canada

Reviewed by:
Henrik Singmann, Albert-Ludwigs-Universität Freiburg, Germany
Gernot D. Kleiter, University of Salzburg, Austria

***Correspondence:**
Jonathan St. B. T. Evans, School of Psychology, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK
j.evans@plymouth.ac.uk;
David E. Over, Department of Psychology, Durham University, South Road, Durham, DH1 3LE, UK
david.over@durham.ac.uk

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *10 December 2014*
Accepted: *20 March 2015*
Published: *08 April 2015*

Citation:
Evans JSBT, Thompson VA and Over DE (2015) Uncertain deduction and conditional reasoning. Front. Psychol. 6:398. doi: 10.3389/fpsyg.2015.00398

Introduction

Paradigm Shift in the Psychology of Reasoning

The psychology of deductive reasoning is undergoing a paradigm shift, which is the consequence of the introduction of Bayesian approaches into the field (see Oaksford and Chater, 2007, 2010; Over, 2009; Manktelow et al., 2011; Elqayam and Over, 2012; Evans, 2012; Baratgin et al., 2013, 2014). In the real world, there are few propositions that people can hold are certainly true, or certainly false, and most of their beliefs come in degrees, which are technically subjective probabilities. We may believe that a grant application has a 50-50 chance of success, or that we will probably be happier if we take a promotion with more responsibility, or that we are unlikely to get on with the new boss we met this morning. It is precisely such uncertain beliefs that we need to take into account when making decisions and solving problems in everyday life. Essential to this process is the ability to combine uncertain beliefs and draw inferences from them, and this is what the new psychology of reasoning is concerned with studying.

The method of study that dominated the field for 40 years or so is the traditional *binary deduction paradigm* (Evans, 2002), inspired by extensional logic and intended to test whether people were capable of logical reasoning without formal training. With this method, participants are given the premises of a logical argument, instructed to *assume that they are true*, and asked to decide whether a purported conclusion *necessarily* follows. They were expected to answer "yes" for arguments considered valid in extensional logic and "no" for those considered invalid. Thus, measured, however, logical reasoning is observed to be generally poor and subject to various cognitive biases (for recent reviews, see Evans, 2007; Manktelow, 2012).

We believe that this traditional paradigm maps quite poorly on to the requirements of real world reasoning. Two key features of the method, which directly reflect the classical binary logic used to assess the accuracy of reasoning, are the instruction to assume the premises and the classification of all statements as simply true or false. High expertise in assumption-based reasoning generally requires specialized training and when logical problems of this kind are administered to naïve participants, we find it unsurprising that error rates are high. We also note that such reasoning loads heavily on working memory and that those of high intelligence do better at these tasks (Evans, 2007; Stanovich, 2011). But everyday reasoning cannot be a specialized tricky business requiring elite professionals and condemning the majority to mistakes. If that were the case, then most people would be incapable of intelligent actions. For these reasons, a number of authors have questioned the relevance of extensional logic and the standard deduction paradigm based upon it (e.g., Oaksford and Chater, 1998, 2007; Evans, 2002; Evans and Over, 2004; Pfeifer and Kleiter, 2010). The new approach treats reasoning as concerning degrees of belief, rather than assumed truth and falsity, and allows that inferences can be drawn with a varying degrees of confidence (Oaksford and Chater, 2007; Evans and Over, 2013).

What is yet to emerge, however, is a clear alternative method for studying reasoning to the traditional deduction paradigm. There are a number of studies which have relaxed instructions, so that participants are given premises but not instructed to assume that they are true, and in which they are sometimes permitted to express degrees of confidence in the conclusions. These are generally known as pragmatic reasoning instructions. Such instructions have been applied to one of the most commonly studied tasks, that of conditional inference. Participants are presented with a conditional and asked whether conclusions follow for four simple inferences, two of which are considered, in most normative systems, as logically valid and two invalid. See **Table 1**.

When the traditional binary paradigm and abstract materials are used (e.g., If the letter is A then the number is 5), participants only show good logical performance on MP, which is nearly always endorsed. MT is also valid but is not endorsed as often as MP, and AC and DA are commonly endorsed, despite being invalid (Evans and Over, 2004). When realistic content is introduced, however, this can substantially affect responding. It has been known for some years that people may resist the simple valid inference MP when they disbelieve the conditional statement (George, 1995; Stevenson and Over, 1995; Politzer, 2005). For example, given the argument

TABLE 1 | The four conditional inferences commonly studied by psychologists.

Modus ponens	MP	If p then q; p therefore q	Valid
Denial of the antecedent	DA	If p then q; not-p therefore not-q	Invalid
Affirmation of the consequent	AC	If p then q; q therefore p	Invalid
Modus tollens	MT	If p then q; not-q therefore not-p	Valid

If the UK builds more nuclear power plants the environment will be safer. (1)
The UK will build more nuclear power plants.
Therefore, the environment will be safer.

Many participants will say that the conclusion does not follow, despite the obvious logic. As the early studies also showed, the exact nature of the instructions is critical. If strict traditional reasoning instructions are employed, with participants asked to assume the premises, they are more likely to resist belief influences and reinstate the inference. However, a recent study has shown the ability to suppress the influence of prior belief on conditional reasoning is restricted to those of higher cognitive ability, even within a university student population (Evans et al., 2010). This difference only occurred under traditional deductive reasoning instructions; with pragmatic reasoning instructions, high ability participants were equally belief "biased." These findings (and many others) suggest to us that assumption-based reasoning is a form of effortful hypothetical thinking (Evans, 2007; Evans and Stanovich, 2013). Belief-based reasoning by contrast is an everyday, natural mode of thought that requires little effort.

If participants are to be allowed to express uncertainty in their conclusions, then are we still studying deduction, or is this a form of inductive inference? In a recent paper, we have shown that deduction in the new paradigm is still distinct from inductive reasoning, but it is described better as what we call *uncertain deduction* (Evans and Over, 2013; see also Pfeifer and Kleiter, 2011). That is, people make deductions in which the uncertainty of the premises is reflected (rightly, according to probability theory) in the uncertainty of the conclusion. Consider a famous piece of reasoning by Sherlock Holmes (see **Table 2**).

Conan Doyle always used the term "deduction," but many readers may have wondered whether the reasoning described is not some type of non-demonstrative inference, such as an abductive inference to the best explanation of the evidence. The conclusions always seem to have a degree of uncertainty (despite being rarely mistaken in the stories). We do not deny that some of Holmes' reasoning is inductive or abductive, and Conan Doyle himself may not have had a very precise understanding of what "deduction" means. But focus on the final sentence above: "Eliminate all other factors, and the one which remains must be the truth." The form of reasoning referred to here is the *disjunctive syllogism*: the logical inference to *q* from the premises *p or q* and *not-p*. Two "factors," *p* and *q*, are referred to in *p or q*, and *not-p* "eliminates" one of these, leaving *q* as what "must" follow. In the story, *p or q* is Watson going to Wigmore Street to send a letter or a wire, and *not-p* is not going there to send a letter, with sending a wire as the conclusion. This inference is clearly deductive, but of course both *p or q* and *not-p* are uncertain to a degree, and the conclusion falls short of certainty. Wigmore Street is just around the corner from Baker Street, and Watson could have gone out for any number of reasons that would have placed him "opposite" the post-office there.

In this example, Holmes' disjunctive syllogism is *classically valid*, in that its conclusion must be true given that its premises are true, but it is not necessarily *sound*. A sound inference is a valid inference the premises of which are actually true. In other

TABLE 2 | Extract from Conan-Doyle's, *The Sign of Four* (1890).

(HOLMES TO WATSON) "Observation shows me that you have been to the Wigmore Street Post-Office this morning, but deduction lets me know that when there you dispatched a telegram."

"Right!" said I. "Right on both points! But I confess that I don't see how you arrived at it. It was a sudden impulse upon my part, and I have mentioned it to no one."

"It is simplicity itself," he remarked, chuckling at my surprise,–"so absurdly simple that an explanation is superfluous; and yet it may serve to define the limits of observation and of deduction. Observation tells me that you have a little reddish mold adhering to your instep. Just opposite the Wigmore Street Office they have taken up the pavement and thrown up some earth which lies in such a way that it is difficult to avoid treading in it in entering. The earth is of this peculiar reddish tint which is found, as far as I know, nowhere else in the neighborhood. So much is observation. The rest is deduction."

"How, then, did you deduce the telegram?"

"Why, of course I knew that you had not written a letter, since I sat opposite to you all morning. I see also in your open desk there that you have a sheet of stamps and a thick bundle of post-cards. What could you go into the post-office for, then, but to send a wire? Eliminate all other factors, and the one which remains must be the truth."

words, we can only be sure of the conclusion if we are sure of the premises. The problem with the classical notion of soundness is that, like classical validity, it is black and white. An argument is either sound or it is not. We might feel some doubt that Holmes' argument is sound, but we are losing something if we totally disregard it. His premises are plausible, and his conclusion is more likely than not. In an uncertain world, that is better than nothing. The new paradigm is really an extension of the old that can deal not just with contexts where statements can be assigned probabilities of 1 ("true") or 0 ("false"), but all values in between. We cannot usually be certain of our premises and conclusions, and have to ask what other degrees of confidence we should have in them. Classical logic does not provide a means for doing this, and we must look elsewhere. The obvious place is in Bayesian subjective probability theory, which extends classical logic in precisely this manner.

Normative Assessment of Uncertain Deduction

The binary and extensional logic of the old deduction paradigm has no means of evaluating inferences from uncertain premises. However, two Bayesian standards, which we have discussed previously (Evans and Over, 2013), can be applied. The first is probabilistic validity, or *p-validity* (Adams, 1998; see also Gilio, 2002; Gilio and Over, 2012). Probabilistic validity is a generalization of classical validity. The latter is truth-preserving. The conclusion of a classically valid inference will be true given that the premises are true: one cannot go from truth in the premises to falsity in the conclusion. Similarly, p-valid inferences are probability-preserving. One cannot go from high probability in the premise of a p-valid single premise inference to low probability in the conclusion. For example, the inference of *and-elimination*, inferring *p* from *p and q*, is p-valid because $P(p \text{ and } q) \leq P(p)$ for all coherent probability assignments. People commit the conjunction fallacy when they violate the p-validity of this inference (Tversky and Kahneman, 1983).

The matter is a bit more complex for inferences with two or more premises. There is a problem of specifying how the probabilities of two or more premises are to be combined, but this is avoided by saying that a p-valid inference cannot take us from low uncertainty in the premises to high uncertainty in the conclusion. We define the *uncertainty* of a proposition *p* as one minus its probability, $1—P(p)$. Then an inference with two or more premises is p-valid if and only if the uncertainty of its conclusion is not greater than the sum of the uncertainties of its premises for all coherent probability assignments. A p-valid deduction from premises cannot increase the uncertainty in the premises; it differs from induction in precisely this respect (Evans and Over, 2013)[1]. In **Table 1**, MP and MT are p-valid inferences, and AC and DA are p-invalid inferences.

To illustrate with conditionals, consider two sets of assignments of probabilities to the premises of an instance of the p-valid inference MP, inferring *q* from the premises *if p then q* and *p*:

	A	B
if p then q	0.8	0.2
p	0.9	0.1

Consider set A first. The sum of the uncertainties of the premises of A is $(1 - 0.8) + (1 - 0.9) = 0.3$. The uncertainty of the conclusion should not exceed that limit, which implies that we would violate p-validity if we assigned a probability to the conclusion *q* of less than 0.7. In that case, we would be in violation of this Bayesian norm by being more uncertain of the conclusion of a p-valid inference than we were of the premises. The formal definition of the p-validity interval for the conclusion probability is shown in **Table 3**. As the uncertainty of the premises increases, the minimum probability value that can be assigned to the conclusion drops. Turning to B, we see that the uncertainties, 0.8 and 0.9, sum to 1.7. Whenever this figure is one or more, it means that we may assign *any* probability between 0 and 1 to the conclusion without violating p-validity. In other words, where premises have low degrees of belief, p-validity can never be violated. This is clearly something that researchers need to take into account. But there is a parallel with the classical position. When we judge that the premises of MP are false, we cannot violate classical validity by holding that the conclusion is also false, because we are not claiming that the conclusion is false when the premises are true.

A further important point about p-validity to stress is that it is defined in terms of coherent probability assignments. For conditional inferences, this coherence depends on the probability of the natural language conditional, *P(if p then q)*. There has been much

[1]A related, but extensional, definition of deduction is that the conclusion cannot convey more semantic information than the premises (Johnson-Laird, 1983).

TABLE 3 | Permitted intervals for conclusions probabilities for the four conditional inferences on two measures.

	p-validity		Coherence	
Inference	**Min**	**Max**	**Min**	**Max**
MP	max{x+y−1,0}	1	xy	1−y+xy
DA	max{x+y−1,0}	1	(1−x)(1−y)	1−x(1−y)
AC	max{x+y−1,0}	1	0	min{y/x,(1−y)/(1−x)}
MT	max{x+y−1,0}	1	max{(1−x−y)/(1−x),(x+y−1)/x}	1

Notes: (1) In each case x = The probability of the major premise, if p then q, and y = the probability of the relevant minor premise, i.e., P(p) for MP, P(not-p) for DA, P(q) for AC, and P(not-q) for MT.

(2) P(if p then q) = P(q|p) is assumed for calculation of the coherence but not p-validity intervals.

(3) For both measures, a "hit" is defined as an estimated conclusion probability which is between the minimum and maximum values shown in the table.

debate in logic, philosophy, and psychology about this probability (Edgington, 1995; Evans and Over, 2004). One possibility is *P(if p then q)* is the probability of the material conditional of elementary extensional logic, *P(not-p or q)*. If this is so, then the assignments $P(if\ p\ then\ q) = P(not\text{-}p\ or\ q) = 0.8$, $P(p) = 0.9$, and $P(q) = 0.7$ are coherent. Another possibility is that *P(if p then q)* is the conditional probability of q given p, $P(q|p)$, and if this is so, $P(if\ p\ then\ q) = P(q|p) = 0.8$, $P(p) = 0.9$, and $P(q) = 0.7$ are incoherent. In fact, making the latter probability judgments is equivalent to the conjunction fallacy, since $P(p\ and\ q) = P(p)(q|p) = 0.72$ and yet $P(q)$ is judged to be 0.7. There are still other possibilities for conditionals based on possible-worlds semantics (Evans and Over, 2004). Nevertheless, judging $P(if\ p\ then\ q) = 0.8$, $P(p) = 0.9$, and $P(q) < 0.7$ is incoherent for all these possible conditionals and violates p-validity, by increasing uncertainty in the conclusion of an inference, MP, which is p-valid for both interpretations of the conditional. To make our study of p-validity as general as possible, and to presuppose as little as possible, we do not make any special assumption about *P(if p then q)* in our study of p-validity. We will simply ask whether people conform to p-validity by making the uncertainty of the conclusion in a conditional inference less than or equal to the sum of the uncertainties of the premises, and whether they conform more to p-validity when they are given explicit inferences. We ask these questions about both the normatively p-valid inferences of MP and MT, and the normatively p-invalid inferences of AC and DA. As we have noted above, people often endorsed AC and DA as "valid" inferences in traditional studies in the binary paradigm, and we wished to test whether they would also do in a probabilistic study.

There are certainly strong arguments (Edgington, 1995) that the probability of the natural language indicative conditional is the conditional probability, that it satisfies what has been called *the Equation*, $P(if\ p\ then\ q) = P(q|p)$. If the Equation holds, the appropriate normative rules for degrees of belief about the natural language conditional are those for conditional probability in Bayesian probability theory. There is much empirical evidence to support the Equation as descriptive of most people's probability judgments (Douven and Verbrugge, 2010; e.g., Evans et al., 2003; Oberauer and Wilhelm, 2003; Over et al., 2007; Politzer et al., 2010; Fugard et al., 2011; Singmann et al., 2014). The majority of participants respect the Equation, but this is by no means universal. It is also found more often in those of high cognitive ability (Evans et al., 2007). The evidence supporting the Equation is at its strongest for the type of realistic conditionals used in our experiment below (see Supplementary Material and Over et al., 2007; Singmann et al., 2014), but we will still not assume that $P(if\ p\ then\ q) = P(q|p)$ in our study of p-validity, for the reason already given.

The second Bayesian standard we will use to assess deduction from uncertain premises is coherence itself. Here our method does presuppose the Equation, $P(if\ p\ then\ q) = P(q|p)$, for otherwise we cannot lay down precise conditions for the coherence of inferences that contain conditionals. We could use "p-consistent" for this generalization of binary consistency (and have done so in Evans and Over, 2013), but p-consistency has been defined in more than one way (Adams, 1998, p. 181), and "coherence" is standard in judgment and decision making. Degrees of belief and subjective probability judgments are coherent when consistent with the axioms of probability theory. Degrees of beliefs in different statements that relate to each other in some way may or may not be coherent. As we saw above, people are incoherent and make judgments equivalent to the conjunction fallacy if they judge that $P(p\ and\ q) > P(p)$. In commenting upon this fallacy, Tversky and Kahneman (1983, p. 313) stated that "…the normative theory of judgment under uncertainty has treated the coherence of belief as the touchstone of human rationality." Their findings have stimulated a rich literature on this fallacy and its possible explanation in terms of the representativeness heuristic (see Tentori et al., 2013, for a recent contribution). Our question in this paper is not whether people are coherent in their conjunction inferences, but rather whether they are coherent in their conditional inferences, and whether their coherence is increased when the conditional inferences are made explicit.

In our approach, $P(q|p)$ is not necessarily given by the ratio, $P(p\ and\ q)/P(p)$, but rather by the *Ramsey test* (Edgington, 1995; Evans and Over, 2004). Using this "test" on *if p then q*, we hypothetically suppose that p holds, while making suppositional changes in our beliefs to preserve consistency, and then make a judgment about q. This procedure allows us to infer a value for $P(q|p)$ when $P(p)$ cannot be fixed because p refers to an action which we are trying to make a decision about, and even when we judge that $P(p) = 0$ (see also Gilio, 2002; Pfeifer and Kleiter, 2009, 2010, 2011; Gilio and Over, 2012).

To illustrate our approach, with the Equation now assumed, suppose we want to make a probability judgment about the

conditional, "If Dr Adler submits her paper to the Journal of Psychology Reports, it will be accepted." We would use the Ramsey test and suppose that she does make the submission, and then using our knowledge of her ability and the standards of the journal, we would make a judgment about the probable acceptance of her paper. Suppose the result is a degree of belief of 0.8 that it will be accepted under that supposition, and with $P(if\ p\ then\ q) = P(q|p)$, our degree of confidence in the conditional will be 0.8. When we take it as certain that Dr Adler will submit her paper to the journal, we should believe 0.80 that it will be accepted, and any other figure would be incoherent. If, however, we have some uncertainty about whether she will submit there or to a journal we have no knowledge of, it becomes more complicated. Suppose we believe only 0.50 that she will submit to the Journal of Psychological Reports and will otherwise submit to the unknown journal. Now we cannot give a specific probability to the paper being accepted, for we lack information about the unknown journal and it acceptance rate.

It is important to understand that in a case like this our belief in the statement "Dr Adler's paper will be accepted" is still constrained. It has to fall within a range of probability values in order to be coherent. Consider the two extreme cases. At one extreme, if Dr Adler submits to the unknown journal, it is certain that the paper will be accepted. So there is a 0.50×0.80 plus a 0.50×1 chance of the paper being accepted, which is 0.90. At the other extreme, it is certain that the paper will be rejected at the unknown journal: now the chance is just $0.50 \times 0.80 = 0.40$ that the paper will be accepted. To be coherent, then, the probability we can set for the paper being accepted has to lie in the interval [0.4, 0.9]. Anything outside of this range is inconsistent with probability theory. **Table 3** shows the formulae for computing this interval for both MP (the case considered) here, and the other three conditional inferences (see Pfeifer and Kleiter, 2009). Note that it is not just the valid inferences that are constrained by coherence. We can compute intervals for all four cases. A study has been reported testing participants for coherence with these equations (Pfeifer and Kleiter, 2010). We also do this but with a different experimental method, as described below.

A probabilistic theory of conditional inference has been presented by Oaksford and Chater (2007; see also Oaksford et al., 2000), and readers may wonder how this relates to the current analysis. These authors consider contexts in which the major premise of a conditional inference is uncertain, but the minor premise is certain. For example, Dr Adler might herself be certain where she will submit her paper. With $P(p) = 1$, the MP interval collapses to $P(q) = P(q|p)$, which is what Oaksford and Chater give as the probability of the conclusion of MP. Their equations for other inferences also take point values for the same reason. Note that participants who conform to Oaksford and Chater's equations will necessarily be in the intervals of **Table 3**. However, we cannot test conformity to their specific equations here, because the minor premises in our materials will rarely be certain (see also Oaksford and Chater, 2013, for an extension of their theory). Indeed, a key purpose of our study is to explore how people take into account the uncertainty in both premises when they reason with conditionals.

The Study

In this study we examine the manner in which naïve participants will assign probabilities to both premises and conclusions of uncertain arguments. In view of the paucity of data on uncertain deduction our principal aim is to discover the extent to which such assignments conform to the two normative standards outlined above: probabilistic validity and coherence with probability theory. Pfeifer and Kleiter (2010) have already reported experiments on uncertain deductions, which they laid out as explicit conditional inferences (see also Singmann et al., 2014). They presented arbitrary premises with explicit probabilities attached and asked participants to indicate the range of probabilities within which the conclusion could fall. These could be compared with the normative equations shown in **Table 3**. They found generally good coherence for MP, but much poorer coherence for the other three inferences.

Our own method differs from of that Pfeifer and Kleiter in several ways. In place of premises with probabilities assigned by the experimenter, we used conditional statements concerning current affairs with evoke real world beliefs (see Supplementary Material). Probabilities were not assigned by the experimenter but taken from the participants themselves. We did this in two different ways. In a Belief group, participants assigned probabilities to the conditionals in one task—*P(if p then q)*—and to the relevant event probabilities in another—that is *P(p)*, *P(not-p)*, *P(q)*, and *P(not-q)*. This is not a reasoning task, of course, and thus can be used to measure what internal consistency, if any, is present in the beliefs expressed. This method has long been used in judgment and decision making, leading most famously to the discovery of the conjunction fallacy (Tversky and Kahneman, 1983) discussed above.

Our second method, more similar to that of Pfeifer and Kleiter (2010), was to lay out the statements as an explicit inference as in the following example:

GIVEN
If more people use sun screen then cases of skin cancer will be reduced
More people will use sun screen
THEREFORE
Cases of skin cancer will be reduced

Participants also assigned probabilities to the three statements here with the inferential structure now clearly cued. We differ from Pfeifer and Kleiter in that our participants provide their own premise probabilities and assign a point value, rather than an interval, to the conclusion. This method allows people to correct incoherence in their belief system as they can now reason explicitly about the way in which uncertainty in the premises should be reflected in the conclusion. We therefore expect stronger conformity to both p-validity and coherence measures in the Inference group.

Method

Participants

Forty six undergraduate students of the University of Saskatchewan participated, with 23 assigned to each of the

two experimental groups; Psychology students received course credits and others were paid for their participation.

Procedure

A set of 48 conditional statements were used concerning real world causal relations, similar to those in previous studies of the authors (Over et al., 2007) run on British participants, and known to vary widely in believability. All concerned causal relations in real world events, such as "If more people use sun screen then cases of skin cancer will be reduced." Where necessary, the sentences were adapted to be relevant to the Canadian context. The sentences used are shown in Supplementary Material. The tasks were administered via computer software with the experimenter present. Participants were instructed that they would be receiving groups of statements which applied to Canada within the next 10 years, and that following each statement they would be asked to indicate the degree to which they believed the statement to be true (expressed as a percentage probability from 0 to 100%). All ratings were provided on a sliding scale located below each of the statements. Participants indicated their responses by clicking a bar on the scale and dragging it to the desired belief percentage. In the Belief group, participants assigned subjective probabilities to a randomized list of the 48 conditionals sentences and separately to a randomized list of the minor premises and conclusions corresponding to each sentence. For example, they gave probabilities for "more people will use sun screen," "more people will not use sun screen," "skin cancer rates will be reduced," and "skin cancer rates will not be reduced" at some point in the list.

In the Inference group, as described above, participant's assigned probabilities to statements grouped as inferences with major premise, minor premise and conclusion rated in immediate succession with the whole argument visible. The headings GIVEN (before the premises) and THEREFORE (before the conclusion were also included). This resulted in another difference between the two groups: those in the inference group rated the same conditional sentence four times in different places as it appeared with each of the inference types, whereas in the belief group, each conditional sentence was rated only once. The order of presentation of each argument was fully randomized so that arguments using the same conditional statement could appear anywhere in the sequence.

Results

As pointed out in the introduction, tests for p-validity are insensitive when the premise probabilities are low. For this reason, all analyses of p-validity reported are for a reduced set of 24 conditional sentences (mean = 60.6, SD = 11.3, Belief group ratings) with substantially higher degrees of belief in the major premise (conditional statement) than the other 24 (mean = 44.5, SD = 13.8). Coherence measures do not suffer from the same problem, so these analyses were conducted using the full set of 48 sentences.

Hit Rates

Our first analyses concern the number of responses considered correct by our two main indices, p-validity and coherence. In each case we can define an interval within which the conclusion probability should be assigned. In case of p-validity, the conclusions have a maximum level of uncertainty, determined by the values actually assigned to the premises. This has to be computed separately for each participant and for each conditional sentence. In the example (A) discussed above, the maximum uncertainty of the conclusion was 0.3, meaning that the minimum probability value for the conclusion was 0.7. We call this value minP. A value of minP was computed from the premises for each participant problem and compared with the value actually assigned by the participant to the conclusion. Any value of minP or above was scored as making a "hit," whether the inference was normatively p-valid, MP and MT, or not, AC and DA (as the participants might consider any of these inferences "valid"). Note that where the maximum uncertainty was 1 or more (as in example B above), minP was set equal to zero. (See **Table 3** for formal definition of the correct interval for the conclusion probability.) A similar approach was used for the coherence analysis, except that here we need to compute two values for the conclusion—minP and maxP—using the equations shown in **Table 3**. Again this target interval depends on the actual probabilities assigned by each participant to each pair of premises. In the coherence analysis, any conclusion probability assigned in the interval [minP, max P] was scored as a hit. (Note that for any given problem minP is computed differently for p-validity than coherence and will not take the same value.)

The frequency of hits for p-validity in the two groups are shown in the white bars of **Figures 1**, **2** (reduced set of higher belief conditionals); an analysis of the chance rates (black bars) is presented in a subsequent section. For the purpose of the ANOVA, we split the four inferences into two factors: Validity (MP, MT vs. DA, AC) and Polarity (MP, AC vs. DA, MT). The main purpose for doing this was to see more clearly whether classically defined valid inferences differed on our measures. In particular, we might expect greater conformity to p-validity on valid inferences, since p-validity is only normatively required for these. The ANOVA revealed several significant findings. As predicted, the Inference group had more hits (mean 0.87) than the Belief group (0.82) [$F_{(1,\ 44)} = 4.27$, MSE = 0.090, $\eta_p^2 = 0.088$, $p < 0.05$]. Contrary to expectations, however, invalid inferences (0.87) had significantly higher p-validity scores than valid inferences (0.83) [$F_{(1,\ 44)} = 9.16$, MSE = 0.064, $\eta_p^2 = 0.172$, $p < 0.005$]. There was also an interaction between the two factors [$F_{(1,\ 44)} = 9.66$, MSE = 0.067, $\eta_p^2 = 0.180$, $p < 0.005$] such that the (reverse) validity effect showed only in the Belief group (compare **Figures 1**, **2**).

We performed an ANOVA for the coherence hit rates (all conditionals) with the same factors—see white bars of **Figures 3**, **4**. There were three significant main effects: Group [$F_{(1,\ 44)} = 17.02$, MSE = 0.567, $\eta_p^2 = 0.279$, $p < 0.001$], as predicted with higher hit rates for the Inference group (0.62) than the Belief group (0.51); Validity [$F_{(1,44)} = 12.88$, MSE = 0.016, $\eta_p^2 = 0.226$, $p < 0.001$]—again higher scores for invalid (0.58) than valid (0.56) inferences, and very large effect of Polarity [$F_{(1,\ 44)} = 52.74$, MSE = 0.363, $\eta_p^2 = 0.545$, $p < 0.001$] reflecting more hits for affirmative (0.66) than negative (0.48) inferences. A Validity by Group interaction [$F_{(1,\ 44)} = 12.88$, MSE = 0.016, $\eta_p^2 = 0.226$, $p < 0.001$] indicated that the (reverse) validity effect was detected

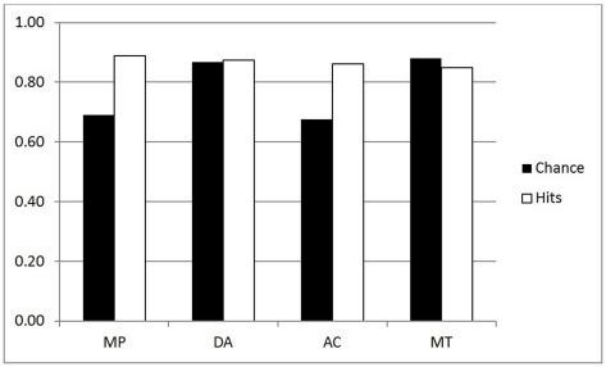

FIGURE 1 | p-validity analysis for the Inference group (Higher belief conditionals).

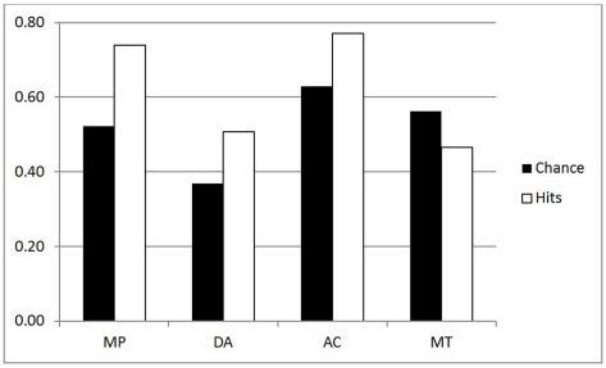

FIGURE 3 | Coherence analysis for the Inference group (all conditionals).

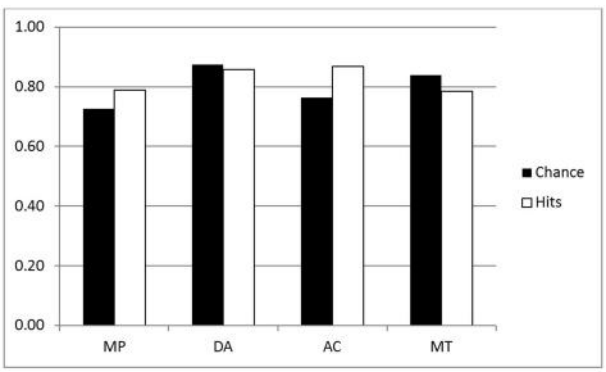

FIGURE 2 | P-validity analysis for the Belief group (Higher belief conditionals).

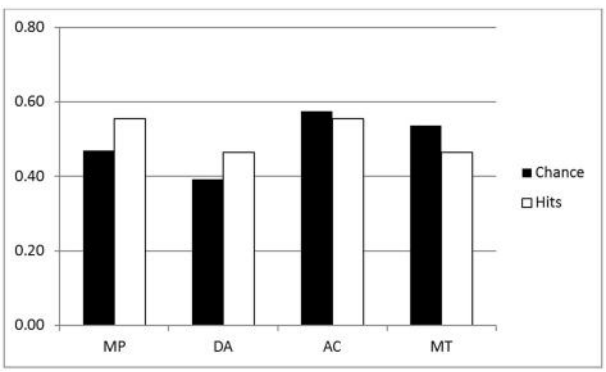

FIGURE 4 | Coherence analysis for the Belief group (all conditionals).

only for the Inference group (the opposite trend to that shown in the p-validity analysis). Finally there was an interaction for Polarity by Group [$F_{(1,\ 44)} = 13.07$, MSE = 0.363, $\eta_p^2 = 0.229$, $p < 0.001$] reflecting a larger effect of Polarity in the Inference than the Belief group.

Before discussing these findings, it is important to consider the chance rates for assigning correct conclusion probabilities which we do next.

Chance Rates

Uncertain deduction presents measurement problems unknown to the standard deduction paradigm. With the old method each conclusion is either valid or not and hence each response either correct or not. With the new method, however, a correct response or "hit" is a value lying within an interval: [minP, 1] for p-validity and [minP, maxP] for coherence. Moreover, these ranges depend upon not only the logical inference under consideration but the actual probabilities assigned to the premises by a particular participant on a particular problem.

The size of these ranges varies considerably and hence the participant has a high chance of guessing the correct answer when they are large. As pointed out in the introduction, where there is low belief in the premises, minP for p-validity may be set to 0, so that *any conclusion probability* will be deemed a hit. For these reasons, it seems essential to consider chance rates and to provide analyses which correct for them[2]. We decided to use the range of the target interval as a measure of chance level responding. For example, with p-validity, if minP was 0.4, we took the value 1-minP = 0.6 to be the chance rate. This is because any participant generating random probabilities with a uniform distribution between 0 and 1, would have a 0.6 chance of hitting the correct interval. For coherence, we took the value (maxP—minP) to be the chance rate for similar reasons. Hence, like hit rates, chance rates have to be computed for each individual participant, conditional and inference. The mean computed chance rates are shown as black bars in **Figures 1–4**.

The first question is whether the observed hit rates were above chance. To assess this, we first computed for each participant the mean difference between hits and chance scores for each conditional sentence, for each inference in both groups on both measures. We then compared these values to a mean of zero with a one sample t test (two tailed, df = 22) in each case. Considering first p-validity, as one might expect from **Figure 1**, scoring was highly significantly above chance for MP and AC in the p-validity

[2]We thank Phil Johnson-Laird for alerting us to this problem.

analysis of the Inference group. Neither DA nor MT were significantly different from chance. For the Belief group (**Figure 2**) hits were again significantly above chance for MP and AC but significantly *below* chance for MT. In the coherence analysis for the Inference group (**Figure 3**) scores were (significantly) above chance for MP, DA, and AC but below chance for MT. In the Belief group (**Figure 4**) all differences were significant with scores above chance for MP and DA and *below* for AC and MT.

Overall, scores were above chance in the majority of cases, but with exceptions. In particular scores for MT tended to be below chance. The high chance rates clearly complicate the interpretation of the analyses of hits reported above. Hence, we decided to repeat these analyses using chance corrected scores, so that the value (hits-chance) was entered as the dependent variable. We refer to these as *performance scores*.

Chance Corrected ANOVAs

Analysis of p-validity

An analysis of variance of was run on the performance scores (hits—chance) for both groups combined on the reduced set of 24 sentences. The factors were Group (Belief vs. Inference), Polarity (MP, AC vs. DA, MT), and Validity (MP, MT vs. AC, DA). All three main effects were statistically significant, the largest being polarity [$F_{(1,\ 44)} = 132.25$, MSE $= 1.198$, $\eta_p^2 = 0.750$, $p < 0.001$], indicating that performance was better on affirmative inferences (MP, AC; mean 0.138) than negative inferences (DA, MT; mean—0.023) as is evident from **Figures 1**, **2** when hits and chance are compared. There was a significant effect of Group [$F_{(1,\ 44)} = 17.95$, MSE $= 0.200$, $\eta_p^2 = 0.290$, $p < 0.001$], showing, as predicted, better performance in the Inference (0.090) than Belief group (0.025). Validity [$F_{(1,\ 44)} = 10.58$, MSE $= 0.031$, $\eta_p^2 = 0.194$, $p < 0.002$] was also significant, as performance was poorer on valid (0.044) than invalid (0.070) inferences due to reversal on MT. There were two significant interactions, one of which was relatively large: Polarity by Group [$F_{(1,\ 44)} = 9.74$, MSE $= 0.088$, $\eta_p^2 = 0.181$, $p < 0.003$]. It is evident from the Figures that the Polarity effect was substantially attenuated in the Belief group. There was also a small but significant three way interaction between Group, Polarity and Validity [$F_{(1,\ 44)} = 4.24$, MSE $= 0.008$, $\eta_p^2 = 0.088$, $p < 0.05$].

Analysis of coherence

Coherence tests apply regardless of the believability of the conditional statement, and so for this measure we report analyses of all 48 sentences. Chance and hit rates are shown on this measure in **Figures 3**, **4** for the Inference and Belief groups respectively. For the Inference group, performance appears to be well above chance for MP, DA, and AC but below chance for MT. Performance appears lower generally in the Belief group but the reverse trend for MT is still present.

The ANOVA for performance scores produced three large effects: Group [$F_{(1,\ 44)} = 16.63$. MSE $= 0.523$, $\eta_p^2 = 0.437$, p $<$ 0.001] with higher scores for Inference (0.091) than Belief (0.016); Polarity [$F_{(1,\ 44)} = 47.87$, MSE $= 0.332$, $\eta_p^2 = 0.521$, $p < 0.001$ with higher scores for MP, AC (0.096)] than for DA, MT (0.011); and Validity by Polarity [$F_{(1,\ 44)} = 167.95$, MSE $= 1.230$, $\eta_p^2 = 0.792$, $p < 0.001$]. The main effect of Polarity and its interaction with Validity reflect the fact that performance reversed on MT for both groups (see **Figures 3**, **4**). Also significant in this analysis were Validity [$F_{(1,\ 44)} = 7.59$, MSE $= 0.033$, $\eta_p^2 = 0.147$, $p < 0.01$], Polarity by Group [$F_{(1,\ 44)} = 18.77$, MSE $= 0.130$, $\eta_p^2 = 0.299$, $p < 0.001$] and Group by Validity by Polarity [$F_{(1,\ 44)} = 8.82$, MSE $= 0.065$. $\eta_p^2 = 0.167$, $p < 0.01$]. The validity effect is also due to poor performance on MT. The three way interaction reflects the fact that the Group by Validity interaction was more marked in the Inference group where performance on inferences other than MT was higher.

Statement Probabilities

As indicated above, chance calculations depend upon the probabilities participants assign to the premises of each argument. Hit rates depend on the conclusion probability assigned. To aid in interpretation of the above findings, we examined the ratings of these statements directly. First, we looked at major premises—the conditional statements themselves. We checked for the Inference group whether conditionals were rated differently on the four occasions they appeared (with each inference). They did not, mean scores being almost identical. We compared the average of these with the single ratings of the same conditionals in the Belief group and they were again similar: Inference 45.2 (SD 19.1), Belief 47.3 (SD 23.4). A t test conducted across the 48 sentences showed no significant difference ($t = 0.62$).

Then we considered the ratings of the events p, not-p, q and not-q which comprise the minor premises and conclusions for the arguments. In the Belief group these are only rated once, but in the Inference group each is rated twice, once when acting as a premise (e.g., p for MP) and once as a conclusion (e.g., p for AC). Ratings as premises and conclusion were again extremely similar in all cases. There were however, substantial differences in the ratings given to affirmative events (p and q) with a mean of 0.52 and for negative events (not-p and not-q) with a mean of 0.39. This effect was very large as shown by an ANOVA [$F_{(1,\ 44)} = 82.53$, MSE $= 0.324$, $\eta_p^2 = 0.652$, $p < 0.0001$]. There was also marginally significant ($p < 0.06$) trend for antecedent events (0.48) to be rated higher than consequent events (0.45).

It is important to note that ratings of affirmative and negative events were incoherent, i.e., inconsistent with probability theory. As **Figure 5** illustrates, the sum of events and their negations was less than one in all cases, whether calculated for the full sets of 48 conditionals or the reduced set of 24. This incoherence has important implications for our findings. In the p-validity analyses (**Figures 1**, **2**) chance rates were significantly higher for DA and MT which make use of negated events. This would follow from underestimation of negative events, because as we have shown earlier, lower belief in premises results in larger ranges for hits on this measure. It also affects chance rates for coherence measures (**Figures 3**, **4**) but in the opposite direction. If assignments were coherent, then we would compute the same chance rates for MP and DA and the same for AC and MT. The former pair use *P(p)* and 1—*P(p)*, which should add to one, the latter *P(q)* and 1—*P(q)* which should also add to one.

To see why underestimating negative event probabilities reduces chance scores for the coherence measure, we take an example. Suppose for a particular conditional a participant sets $P(q|p) = 0.7$, $P(p) = 0.6$ and $P(not\text{-}p) = 0.32$. This shows the typical bias in our experiment, as $P(p) + P(not\text{-}p) = 0.92$ overall.

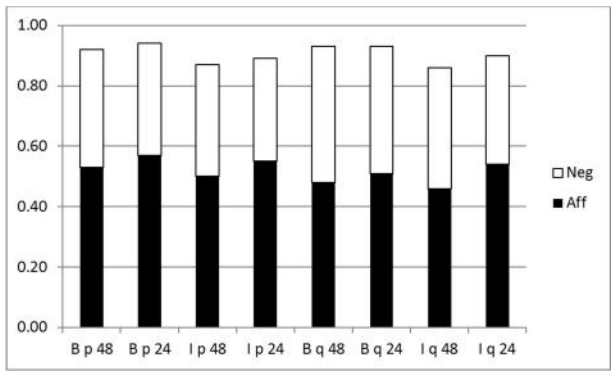

FIGURE 5 | Stacked bar chart showing probabilities assigned to events and their negations. B, Belief group; I, Inference group; 48, full set of conditionals; 24, reduced set; p, antecedent event (black bar p, white bar not-p); q, consequent event (black bar q, white bar not-q).

When we compute the hit interval for MP, by the equations given in **Table 1**, we get [0.42, 0.82] with a chance calculation of 0.40. Had *P(not-p)* been assigned coherently, i.e., to 0.4, the interval for DA would compute to be [0.12, 0.58] again with a chance rate of 0.40. However, it is underestimated, resulting in a computed interval of [0.476, 0.796] and a chance rate of 0.32 which is lower than is should be.

Discussion

The objective of the present study was to investigate the accuracy with which people can make probability judgments about the premises and conclusions of conditional inferences, and to test whether this accuracy, as measured by Bayesian standards, is increased by explicit conditional reasoning. We appealed to the standards of p-validity and coherence. We used two methods: a Belief group who rated beliefs in the statements presented separately, and an Inference group who saw them grouped as an explicit inference. We found that our participants did conform to p-validity at rates significantly higher than chance, but only for the affirmative inferences MP and AC. This performance was also significantly higher for the Inference group. Results were similar for the coherence measure. Again performance was well above chance for MP and AC, and significantly better for the Inference group. However, the results for the denial inference DA and MT were more complex, as participants were above chance for the former and below chance for the latter.

As must be evident to the reader, the study of uncertain deduction is a good deal more complex than use of the traditional deduction paradigm. In the traditional method, each inference is classified as valid or invalid and the participant either does or does not endorse the inference. To study uncertain deduction we must allow participants to assign probabilities to the premises and the conclusions of deductive arguments. The difficulty then comes in assessing whether they have done this correctly. First, there is not one but two different measures that can be taken: p-validity and coherence. Second, each of these allows participants to assign a conclusion probability within an interval. This interval must be computed for each participant on each problem separately depending on the premise probabilities assigned. Finally, these intervals can be large, creating the problem that participants may hit them by chance. We have shown in this paper how to compute these chance intervals and proposed method to correct hits rates for guessing.

Very little previous work has been conducted on uncertain deduction, despite apparent enthusiasm for a new paradigm psychology of reasoning based on degrees of belief rather than black and white truth judgments. The methodology introduced here differs in significant ways from the study of Pfeifer and Kleiter (2009, 2010) who studied only coherence (not p-validity), using premise probabilities assigned by the experimenter and allowing participants to assign a range of probabilities to the conclusion. Their results differ from ours in that they found coherence to be good only for MP, whereas we find this to be the case for MP, AC, and DA. This could reflect the difference in response method, but we think it more likely due to our use of realistic, causal-temporal conditional statements which introduce real world experience of causal relations. (We have no account of the reversal on MT, however.) In addition to assessing the coherence of conclusion probabilities taken as point ratings, we believe this to be the first psychological study to measure directly whether people conform to p-validity when both major and minor premises are taken to be uncertain. In both cases, this means that a range of values are acceptable as a "hit" on either measure. We consider our two measures in turn.

Probabilistic validity, or p-validity, is a relatively weak measure for us. For generality and to minimize our assumptions, we did not presuppose that $P(\text{if } p \text{ then } q) = P(q|p)$ in our assessment of p-validity, but simply assessed whether participants express no more uncertainty in the conclusion than in the premises of our conditional inferences. This notion of validity does not constrain conclusion probabilities for the invalid inferences, AC and DA, nor in effect, for valid inferences with low belief premises. Hit rates generally exceeded chance in our study only for the affirmative inferences MP and AC. Chance rates are disturbingly high (black bars, **Figures 1**, **2**) even with the analysis restricted to the higher belief conditionals. Hence, we suggest that this measure will only be useful for problems where there is a very high degree of belief in the premises. Nevertheless, we have some findings of interest on this measure. First, as predicted, p-validity scores are higher for the Inference than the Belief group, with and without chance correction. The second finding of particular interest is that participants did not conform more to p-validity on the inferences that are actually valid, MP and MT. Indeed there was a small trend in the opposite direction. Much larger was an effect of polarity such that participants performed better on the affirmative inferences, MP and AC.

These findings can be accounted for as follows. First, the chance rates are very high on DA and MT due to underestimation of negative event probabilities, as explained earlier. This creates a ceiling effect for these two inferences, making it difficult for participants to perform above chance. This does not explain, however, why performance is equally high on MP and AC and facilitated for the Inference group in both cases. Research in the traditional paradigm often showed high endorsement of

AC, though it is both classically invalid and p-invalid (Evans and Over, 2004). It may be that the participants interpreted the assertion of our causal-temporal conditionals (in Supplementary Material) *if p then q* as also pragmatically implying *if q then p*, making AC in effect MP in the other direction. That could explain their apparently equal effort to generate a p-valid conclusion in the Inference group for AC as for MP. In a study corresponding to our Inference group, Singmann et al. (2014) assessed p-validity only for MP and MT, and found that participants conformed to p-validity for MP and not MT. Still, as we have explained above, pragmatic factors can have a large effect people's reasoning. There could be pragmatic differences in the materials used by Singmann et al. and ourselves, and further research must investigate this possibility.

It could also be suggested that our use of causal-temporal conditionals, *if p then q*, implies not only that *P(q|p)* is high but that *P(q|not-p)* is low, in conformity with the *delta-p rule*, *P(q|p)—P(q|not-p)*, which measures how far that *p* raises the probability of *q*. It is true that, when *p* causes *q*, *p* would normally be thought to raise the probability of *q*, but previous work has not found that people interpret causal-temporal conditionals in terms of the delta-p rule (see Over et al., 2007, and especially Singmann et al., 2014, on this rule).

Use of the coherence measures allows us to ask whether people are coherent in the beliefs they express about conditional statements and their component events. This measure is stronger than that of p-validity and is applicable to both p-valid and p-invalid inferences. But the equations we use for coherence do assume that *P(if p then q) = P(q|p)*, which, as we explained above, is often called the Equation (Edgington, 1995). Examining the data, we have found again that coherence is better for the Inference than the Belief group, again with and without chance correction. As with p-validity, these analyses are affected by the underestimation of negative event probabilities, which in this case causes chance rates to drop somewhat for DA and MT. But it is striking that the facilitation of coherence in the Inference group is restricted to MP, DA, and AC, as can be seen by comparing **Figures 3**, **4** (see yet again Singmann et al., 2014, and recall our point about possible pragmatic differences between their materials and ours).

Interpretation of findings on the negative inferences, DA and MT, is clearly complicated by the underestimation of negative event probabilities we have observed. If we focus our attention on the affirmative inferences, MP and AC, however, it is clear that participants perform well above chance on both measures in the Inference group. In other words when given the opportunity to see the statements grouped as an inference, untrained participants do seem to grasp intuitively the logical restrictions that premise probabilities place upon conclusion probabilities. The actual hit rates are well over 80% for p-validity and around 75% for coherence. We find these figures quite encouraging, as supporting the conclusion that one way to improve Bayesian reasoning is by the use of explicit inferences. Explicit reasoning may not always make people rational by Bayesian standards, but it can help (see also Cruz et al., 2015).

Uncertain deduction is central to the new paradigm psychology of reasoning. If research is to progress, we must find methods for studying the relation between belief in premises and belief in conclusions. It is, as we have shown, a much trickier task than that presented by the standard deduction paradigm. There are a number of pointers to future research studies arising from our findings. For example, studies of p-validity should be restricted to problems with high belief (but still uncertain) premises, in order to provide sufficient sensitivity. We have also highlighted a problem with explicitly negated premises. Events expressed as negations tend to be underestimated in their probabilities, providing an immediate source of incoherence. This could be related to the findings in "support theory" of *subaddivity*: that the weight given to an implicit disjunction is less than the sum of its disjuncts when these are made explicit (Tversky and Koehler, 1994). A negated event is itself an implicit disjunction; that is, not-A consists of B v C v …, which are the explicit alternatives. For example, the probability assigned to "school class sizes are not reduced" might be less than the sum that would be assigned to "school class sizes are increased" and "school class sizes remain the same." In any event, this problem must be addressed in future studies of the coherence of negated inferences[3].

We believe that there is much to be gained from the further study of the coherence of conditional beliefs, as in our Belief group. We have noted above the rich literature that resulted from the discovery of the conjunction fallacy. The representativeness heuristic that Tversky and Kahneman (1983) proposed as an explanation of this incoherence in conjunctive beliefs might also cause some incoherence in conditional beliefs, but other, as yet unknown heuristics could play a role as well. We hope to have demonstrated here, however, that the study of deductive reasoning using Bayesian methods should move beyond the almost exclusive focus on the inference from *p and q* to *q* and the associated conjunction fallacy. There is much more to discover about Bayesian reasoning by studying other deductive inferences with uncertain premises.

In an ideal Bayesian world, probabilities assigned to logically related statements would be perfectly coherent with probability theory, but in reality this is unlikely to hold, especially when the statements are not explicitly related as inferences. Such probabilities are unlikely to be assigned on an absolute basis due to the power of pragmatics in human communication and understanding. We interpret statements in their context, amplifying their meanings and making probability judgments with implicit heuristics. It is unsurprising that people's beliefs are not fully coherent. It is impossible for them to ensure absolute coherence, even in relatively simple beliefs, due computationally intractability. However, it is of great interest to discover the causes of incoherence in conditional beliefs, such as the difficulty with negative events reported here.

Grouping uncertain statements together as an inference is a natural way to extend the traditional deduction paradigm to the study of uncertain deduction. The fact that participants in our Inference group consistently performed better than participants in the Belief group might suggest that the former were intervening with explicit reasoning in order the make their judgments

[3]The coherence intervals of **Table 3** are derived using the total probability theorem of probability theory. See Hadjichristidis et al. (2014) on this theorem and their findings of superaddivity.

more consistent. Further research will be needed, however, to determine whether this is in fact that case. An alternative pragmatic account is that concurrent presentation of premises and conclusions contextualizes the statements together so that judgments become more consistent without any conscious effort of reasoning. If explicit reasoning is involved, this could be indicated by examining performance under working memory load or by correlating performance with individual measures of cognitive ability. These are among the methods employed by dual process researchers to identify effortful reasoning (Evans and Stanovich, 2013).

In conclusion, we hope to have developed a methodology that can be adapted for a variety of future uses in the new psychology of deduction. We have shown that it is feasible to study the relation between the degree of belief that people hold in premises and conclusion of a logical argument. We have also shown that such judgments are not random and conform to the coherence of probability theory at rates well above that which could be expected by chance. People have some problems with the coherence of their judgments about negative events, but are otherwise fairly good, by Bayesian standards, at conditional reasoning. Their performance is at an even higher level when statements are grouped together into explicit conditional inferences.

Acknowledgments

The authors would like to thank Christoph Klauer, Gernot Kleiter, Niki Pfeifer, and Henrik Singmann for much helpful discussion of issues closely related to the topic of this paper and Nicole Therriault for her help in gathering the data. We also thank Mike Oaksford and Phil Johnson-Laird for their comments on an earlier version of this manuscript. Support for this project was provided by and Natural Engineering and Research Council Discovery Grant to the second author.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2015.00398/abstract

References

Adams, E. (1998). *A Primer of Probability Logic*. Stanford, CA: CLSI publications.

Baratgin, J., Over, D. E., and Politzer, G. (2013). Uncertainty and the de Finetti tables. *Think. Reason.* 19, 308–328. doi: 10.1080/13546783.2013.809018

Baratgin, J., Over, D. E., and Politzer, G. (2014). New psychological paradigm for conditionals: its philosophy and psychology. *Mind Lang.* 29, 73–84. doi: 10.1111/mila.12042

Cruz, N., Baratgin, J., Oaksford, M., and Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Front. Psychol.* 6:192. doi: 10.3389/fpsyg.2015.00192

Douven, I., and Verbrugge, S. (2010). The Adams family. *Cognition* 117, 302–318. doi: 10.1016/j.cognition.2010.08.015

Edgington, D. (1995). On conditionals. *Mind* 104, 235–329. doi: 10.1093/mind/104.414.235

Elqayam, S., and Over, D. E. (2012). New paradigm in psychology of reasoning: probabilities, beliefs and dual processing. *Mind Soc.* 11, 27–40. doi: 10.1007/s11299-012-0102-4

Evans, J. St. B. T. (2002). Logic and human reasoning: an assessment of the deduction paradigm. *Psychol. Bull.* 128, 978–996. doi: 10.1037/0033-2909.128.6.978

Evans, J. St. B. T. (2007). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. Hove: Psychology Press.

Evans, J. St. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Think. Reason.* 18, 5–31. doi: 10.1080/13546783.2011.637674

Evans, J. St. B. T., Handley, S. J., and Over, D. E. (2003). Conditionals and conditional probability. *J. Exp. Psychol.* 29, 321–335. doi: 10.1037/0278-7393.29.2.321

Evans, J. St. B. T., Handley, S., Neilens, H., Bacon, A. M., and Over, D. E. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *Q. J. Exp. Psychol.* 63, 892–909. doi: 10.1080/17470210903111821

Evans, J. St. B. T., Handley, S., Neilens, H., and Over, D. E. (2007). Thinking about conditionals: a study of individual differences. *Mem. Cogn.* 35, 1772–1784. doi: 10.3758/BF03193509

Evans, J. St. B. T., and Over, D. E. (2004). *If*. Oxford: Oxford University Press.

Evans, J. St. B. T., and Over, D. E. (2013). Reasoning to and from belief: deduction and induction are still distinct. *Think. Reason.* 19, 267–283. doi: 10.1080/13546783.2012.745450

Evans, J. St. B. T., and Stanovich, K. E. (2013). Dual process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685

Fugard, J. B., Pfeifer, N., Mayerhofer, B., and Kleiter, G. D. (2011). How people interpret conditionals: shifts toward conditional event. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 635–648. doi: 10.1037/a0022329

George, C. (1995). The endorsement of the premises: assumption-based or belief-based reasoning. *Br. J. Psychol.* 86, 93–111. doi: 10.1111/j.2044-8295.1995.tb02548.x

Gilio, A. (2002). Probabilistic reasoning under coherence in System P. *Ann. Math. Artif. Intell.* 34, 5–34. doi: 10.1023/A:1014422615720

Gilio, A., and Over, D. E. (2012). The psychology of inferring conditionals from disjunctions: a probabilistic study. *J. Math. Psychol.* 56, 118–131. doi: 10.1016/j.jmp.2012.02.006

Hadjichristidis, C., Sloman, S. A., and Over, D. E. (2014). Categorical induction from uncertain premises: jeffrey deson't completely rule. *Think. Reason.* 20, 405–431. doi: 10.1080/13546783.2014.884510

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge: Cambridge University Press.

Manktelow, K. I. (2012). *Thinking and Reasoning*. Hove: Psychology Press.

Manktelow, K. I., Over, D. E., and Elqayam, S. (2011). "Paradigm shift: jonathan evans and the science of reason," in *The Science of Reason: A Festchrift for Jonathan St B T Evans,* eds K. I. Manktelow, D. E. Over, and S. Elqayam (Hove: Psychology Press), 1–16.

Oaksford, M., and Chater, N. (1998). *Rationality in an Uncertain World*. Hove: Psychology Press.

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Oaksford, M., and Chater, N. (2010). "Cognition and conditionals: an introduction," in *Cognition and Conditionals: Probability and Logic in Human Thinking,* eds M. Oaksford and N. Chater (Oxford: Oxford University Press), 3–36.

Oaksford, M., and Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Think. Reason.* 19, 346–379. doi: 10.1080/13546783.2013.808163

Oaksford, M., Chater, N., and Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *J. Exp. Psychol.* 26, 883–889. doi: 10.1037//0278-7393.26.4.883

Oberauer, K., and Wilhelm, O. (2003). The meaning(s) of conditionals: conditional probabilities, mental models and personal utilities. *J. Exp. Psychol.* 29, 680–693. doi: 10.1037/0278-7393.29.4.680

Over, D. E. (2009). New paradigm psychology of reasoning. *Think. Reason.* 15, 431–438. doi: 10.1080/13546780903266188

Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., and Sloman, S. A. (2007). The probability of causal conditionals. *Cogn. Psychol.* 54, 62–97. doi: 10.1016/j.cogpsych.2006.05.002

Pfeifer, N., and Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *J. Appl. Logic* 7, 206–217. doi: 10.1016/j.jal.2007.11.005

Pfeifer, N., and Kleiter, G. D. (2010). "The conditional in mental probability logic," in *Cognition and Conditionals: Probability and Logic in Human Thinking,* eds M. Oaksford and N. Chater (Oxford: Oxford University Press), 153–173.

Pfeifer, N., and Kleiter, G. D. (2011). "Uncertain deductive reasoning," in *The Science of Reason: A Festchrift for Jonathan St B T Evans,* eds K. I. Manktelow, D. E. Over, and S. Elqayam (Hove: Psychology Press), 145–166.

Politzer, G. (2005). Uncertainty and the suppression of inferences. *Think. Reason.* 11, 5–33. doi: 10.1080/13546780442000088

Politzer, G., Over, D. E., and Baratgin, J. (2010). Betting on conditionals. *Think. Reason.* 16, 172–197. doi: 10.1080/13546783.2010.504581

Singmann, H., Klauer, K. C., and Over, D. E. (2014). New normative standards of conditional reasoning and the dual-source model. *Front. Psychol.* 5:316. doi: 10.3389/fpsyg.2014.00316

Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. New York, NY: Oxford University Press.

Stevenson, R. J., and Over, D. E. (1995). Deduction from uncertain premises. *Q. J. Exp. Psychol.* 48A, 613–643. doi: 10.1080/14640749508401408

Tentori, K., Crupi, V., and Russo, S. (2013). On the determinants of the conjunction fallacy: probability vs inductive confirmation. *J. Exp. Psychol.* 142, 235–255. doi: 10.1037/a0028770

Tversky, A., and Kahneman, D. (1983). Extensional vs intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90, 293–315. doi: 10.1037/0033-295X.90.4.293

Tversky, A., and Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychol. Rev.* 101, 547–567. doi: 10.1037/0033-295X.101.4.547

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 25 February 2015
doi: 10.3389/fpsyg.2015.00192

Bayesian reasoning with ifs and ands and ors

Nicole Cruz[1], Jean Baratgin[2,3], Mike Oaksford[1] and David E. Over[4]*

[1] Department of Psychological Sciences, Birkbeck, University of London, London, UK
[2] Laboratory CHArt (PARIS), Université Paris 8, Paris, France
[3] Institut Jean Nicod, Paris, France
[4] Department of Psychology, Durham University, Durham, UK

Edited by:
David R. Mandel, Defence Research and Development Canada, Toronto Research Centre, Canada

Reviewed by:
Niki Pfeifer, Ludwig-Maximilians-Universität München, Germany
Igor Douven, University of Groningen, Netherlands

***Correspondence:**
David E. Over, Department of Psychology, Durham University, Durham University Science Site, South Road, Durham DH1 3LE, UK
e-mail: david.over@durham.ac.uk

The Bayesian approach to the psychology of reasoning generalizes binary logic, extending the binary concept of consistency to that of coherence, and allowing the study of deductive reasoning from uncertain premises. Studies in judgment and decision making have found that people's probability judgments can fail to be coherent. We investigated people's coherence further for judgments about conjunctions, disjunctions and conditionals, and asked whether their coherence would increase when they were given the explicit task of drawing inferences. Participants gave confidence judgments about a list of separate statements (the statements group) or the statements grouped as explicit inferences (the inferences group). Their responses were generally coherent at above chance levels for all the inferences investigated, regardless of the presence of an explicit inference task. An exception was that they were incoherent in the context known to cause the conjunction fallacy, and remained so even when they were given an explicit inference. The participants were coherent under the assumption that they interpreted the natural language conditional as it is represented in Bayesian accounts of conditional reasoning, but they were incoherent under the assumption that they interpreted the natural language conditional as the material conditional of elementary binary logic. Our results provide further support for the descriptive adequacy of Bayesian reasoning principles in the study of deduction under uncertainty.

Keywords: uncertain reasoning, deduction, conditionals, coherence, conjunction fallacy

INTRODUCTION

Most everyday and scientific inferences are from uncertain premises, with the aim of forming and revising beliefs and making decisions. For example, some hypotheses about global warming are more highly confirmed than others, but all are uncertain to some degree, and yet there have to be inferences from these hypotheses to further scientific research and practical decision making. Given the ubiquity of reasoning under uncertainty, an important question in the psychology of reasoning is how good people are at it, and what can improve it when it falls short of the appropriate normative theory.

Tversky and Kahneman (1983) pointed out that "...the normative theory of judgment under uncertainty has treated the coherence of belief as the touchstone of human rationality." *Coherence* is the normative foundation of the Bayesian approach to the study of cognition (Chater and Oaksford, 2008), which is having an immense impact on the psychology of reasoning (Elqayam and Over, 2013). To be coherent is to conform to the axioms of probability theory, which are justified by the Dutch book theorem (de Finetti, 1974).

There are tasks and contexts in which there appears to be a remarkably good correspondence between people's probability judgments and probability theory (Griffiths and Tenenbaum, 2006; Oaksford and Hahn, 2007; Fiser et al., 2010; Oaksford and Chater, 2013). But there are also contexts in which people are incoherent. Until very recently, there were only limited studies of whether people are coherent in their judgments about the basic logical connectives of conjunction, disjunction, and the conditional. Of course, there have been innumerable papers on the *conjunction fallacy* (Tversky and Kahneman, 1983): judging that the probability of a conjunction, *P(p and q)*, is greater than the probability of one of its conjunctions, *P(p)*. The valid logical inference related to this fallacy is *and-elimination*: inferring *p* from *p and q*. But this is just one out of many logical inferences in which conjunction occurs. There have been relatively few studies of the *disjunction fallacy* (Bar-Hillel and Neter, 1993): judging that *P(p)* is greater than *P(p or q)*. The valid inference for this fallacy is *or-introduction*: inferring *p or q* from *p*. There should be wider studies of probability judgments about conjunctions and disjunctions, especially when these connectives are related to the conditional, *if p then q*, since conditionals are at the heart of so much reasoning in both everyday affairs and science.

The purpose of this paper is to extend the study of whether people's probability judgments about conjunctions, disjunctions, and conditionals are coherent. Our approach is that of the new paradigm in the psychology of deductive reasoning, which goes beyond the binary distinction between categorical belief in the truth, or falsity, of propositions to the full range of degrees of belief, or subjective probabilities (Evans and Over, 2004, 2013; Oaksford and Chater, 2007, 2012, 2013; Pfeifer and Kleiter, 2010,

2011; Baratgin et al., 2013; Pfeifer, 2013; Over, in press). The probabilistic approach has taken two important steps in the study of deduction: (1) it represents uncertainty in the premises and conclusions of inferences, and (2) it represents the probability of the natural language indicative conditional, *P(if p then q)*, as the conditional probability of *q* given *p*, *P(q|p)*. The relation, *P(if p then q) = P(q|p)*, is so fundamental for a Bayesian account of conditional reasoning that it has simply been called *the Equation* (Edgington, 1995; Oaksford and Chater, 2007). A conditional that satisfies the Equation has been called the *probability conditional* (Adams, 1998; Oaksford and Chater, 2007), but we call it here the *conditional event* (following de Finetti, 1995). The conditional probability in the Equation, *P(q|p)*, is not defined by the ratio *P(p and q)/P(p)* in our approach (see also Pfeifer, 2014). One can easily think of cases in which people have a clear degree of belief about *P(q|p)* even though they judge that *P(p) = 0*, or they cannot make a judgment at all about *P(p)* (Adams, 1998). We rather argue that people infer the conditional probability in a *Ramsey test*, that is, a mental simulation in which they hypothetically suppose *p* to be the case, make whatever changes to their beliefs are necessary to preserve consistency, and assess the probability of *q* on this basis (Stalnaker, 1968; Ramsey, 1994; Evans and Over, 2004). Both (1) and (2) have received strong and converging empirical support (Oaksford et al., 2000; Evans et al., 2003; Oberauer and Wilhelm, 2003; Oberauer et al., 2007; Over et al., 2007; Douven and Verbrugge, 2010; Politzer et al., 2010; Fugard et al., 2011b; Baratgin et al., 2014; Cruz and Oberauer, 2014; Singmann et al., 2014).

In an influential alternative approach, mental model theory, the natural language indicative conditional is taken to have the same full models as the *material conditional* of elementary extensional logic, which is logically equivalent to *not-p or q*. The material conditional is truth functional, that is, its truth or falsity is a function of the truth or falsity of its elementary components, the propositions *p* and *q*. It is false when its antecedent *p* is true and the consequent *q* is false, and it is true in the other three possible cases (that is, the cases *p and q*, *not-p and q*, and *not-p and not-q*). In mental model theory, *P(if p then q) = P(not-p or q)* is supposedly the correct normative probability judgment to make (Johnson-Laird and Byrne, 1991, 2002; Byrne and Johnson-Laird, 2009, 2010). Whether *P(if p then q)* equals *P(q|p)* or *P(not-p or q)* very much affects which judgments are coherent in conditional reasoning, as we will see below.

Consider the uncertain premises and possible uncertain conclusions that form the basis of most of our ordinary and scientific reasoning. The axioms of probability theory can be used to determine whether combinations of these premises and conclusions are, or are not, coherent (for recent examples see Pfeifer and Kleiter, 2005, 2009; Gilio and Over, 2012). For instance, there are the valid inferences of *and-elimination*, referred to above, and also *and-introduction*: inferring *p and q* from the separate premises *p* and *q*. For probability judgments about *p*, *q*, and *p and q* to be coherent, *P(p and q)* must lie in the interval between P(p) + P(q) − 1 (or 0 if this sum is negative) at the lower end, and the minimum of *P(p)* and *P(q)* at the upper end (Pfeifer and Kleiter, 2005). For example, *P(p) = P(q) = 0.6* and *P(p and q) = 0.1* is incoherent because *P(p and q)* is too low, and *P(p) = P(q) = 0.6* and *P(p and q) = 0.7* is incoherent, and the conjunction fallacy is committed, because now *P(p and q)* is too high. Our question in this paper is whether people are generally coherent in their conjunctive, disjunctive, and conditional premises and conclusions, and whether their coherence is improved when they are given explicit inferences. Tversky and Kahneman (1983) did not ask their participants to infer degrees of confidence in the conclusion *p* from an uncertain *p and q* premise in an explicit inference, but we did ask participants in our experiments.

Studies of the coherence between premises and conclusions of people's reasoning has only just begun (Pfeifer and Kleiter, 2005, 2010, 2011; Pfeifer, 2013; Politzer and Baratgin, under review; Singmann et al., 2014; Evans and Over, under review). There is evidence, for example, that people are coherent in explicit *and-introduction* inferences (Pfeifer and Kleiter, 2005; Politzer and Baratgin, under review). There are also some studies of the classical conditional inferences of modus ponens (MP), modus tollens (MT), affirmation of the consequent (AC), and denial of the antecedent (DA), and it has been found that the degree to which people are coherent can increase when they are given some of these conditional inferences as explicit tasks to perform (Evans and Over, under review; see also Pfeifer and Kleiter, 2010; Singmann et al., 2014).

We conducted two experiments focusing on conjunctions, disjunctions, and their relationships with conditionals, and comparing probability judgments about the premises and conclusions when these were given as separate statements and when they were arranged as explicit inferences. Experiment 1 looked at inferences between disjunctions and conditionals, and Experiment 2 at inferences between conjunctions and conditionals. The inferences are summarized in **Table 1**.

Inferences 1.1 and 1.2 are logically equivalent, as are inferences 1.3 and 1.4, as well as inferences 1.5 and 1.6. They differ only in the position of the negation they contain. The two positions of the negation instantiated in the inferences are those for which the largest negation effects have been reported in the literature (Oberauer et al., 2011; Espino and Byrne, 2013). We introduced this variation in order to control for negation effects. Experiment 1 assessed two further inferences, and Experiment 2 six further inferences, which are not listed in **Table 1**. These additional inferences were used to investigate other questions, and are not discussed here further.

Inferences 1.1 and 1.2 are logically equivalent forms of *or-introduction*, and here it is clearly incoherent to judge that the probability of the conclusion is lower than that of the premise. It is a consequence of the axioms of probability theory that *P(p) ≤ P(p or q)*. In the binary approach, it is inconsistent to assume the

Table 1 | The inferences used in Experiments 1 and 2.

	Experiment 1		Experiment 2
1.1	p, therefore p or q	2.1	p & q, therefore if p then q
1.2	not-p, therefore not-p or q	2.2	p, q, therefore if p then q
1.3	If p then q, therefore not-p or q	2.3	p & q, therefore p
1.4	if not-p then q, therefore p or q	2.4	p & q, therefore q
1.5	p or q, therefore if not-p then q		
1.6	not-p or q, therefore if p then q		

truth of the premise of *or-introduction*, *p*, but not to accept that *p or q* follows. Binary studies found that people did endorse this inference at just above chance levels, but also that there was significant resistance to it (Rips, 1983, 1994; Braine et al., 1984). This finding has generally been explained as a pragmatic effect: people are unwilling to draw the inference because it would be misleading in a conversation with another person to endorse *p or q* when one can make the more informative statement *p* (Grice, 1989; see also Bar-Hillel and Neter, 1993; Tversky and Koehler, 1994; Fugard et al., 2011a). The much wider Bayesian approach can cover the special case of binary inconsistency by letting a probability of 1 represent "true" and a probability 0 represent "false." The binary findings could be said to reveal implicit incoherent reasoning because people are, in effect, making $P(p) = 1$, *p* is "true," and $P(p \text{ or } q) = 0$, *p or q* is "false." However, we predicted greater coherence when people are explicitly asked for their degrees of belief about *p* and *p or q*. People can then state their degrees of belief directly, without needing to consider additional pragmatic factors that arise when communicating with another speaker. This should lead to the prevention, or at least to a strong reduction, of pragmatic effects.

Inferences 1.3 and 1.4 are logically equivalent *if-to-or* inferences and go from a conditional to a disjunction. Supposing *if p then q* is equivalent to the material conditional, $P(\text{if } p \text{ then } q) = P(\text{not-}p \text{ or } q)$, any other judgment is incoherent. Supposing *if p then q* is the conditional event, $P(\text{if } p \text{ then } q) = P(q|p)$. It follows from the axioms of probability theory that $P(q|p) \leq P(\text{not-}p \text{ or } q)$, and probability judgments must conform to this relation to be coherent.

Inferences 1.5 and 1.6 are logically equivalent *or-to-if* inferences and go from a disjunction to a conditional. If the conditional in these inferences is interpreted as the material conditional, then the same equivalence holds as for 1.3 and 1.4, and judgments are only coherent when the premise and conclusion are assigned the same probability. If the conditional is interpreted as the conditional event, then judgments are coherent when they conform to the relation $P(q|p) \leq P(\text{not-}p \text{ or } q)$. Thus the relation that must hold for the inferences to be coherent is the same for 1.3–1.4 and for 1.5–1.6. The difference is that in the first two the conditional is the premise, and in the second two the conditional is the conclusion. This implies that, if one interprets the conditional as the conditional event, the *if-to-or* inference is coherent when the probability of the conclusion is equal or higher than that of the premise, whereas the *or-to-if* inference is coherent when the probability of the conclusion is equal or lower than that of the premise. This difference in the conditions for coherence of the two inferences is reflected in the fact that under a conditional event interpretation, the *if-to-or* inference is valid, whereas the *or-to-if* inference is invalid and can even be a quite a weak inference.

When we speak of validity in this context of uncertain inference, we mean probabilistic validity, or *p-validity*. P-validity is a generalization of binary validity to reasoning under uncertainty. Just as an inference is binary valid when there are no cases in which the conclusion is false and the premise is true, a single premise inference is p-valid when there are no coherent cases in which the probability of the conclusion is lower than the probability of the premise (see Adams, 1998, on p-validity for inferences with more than one premise; Singmann et al., 2014; Evans and Over, under review, for applications in the psychology of reasoning). For the *or-to-if* inference, such cases are possible. Consider an instance of 1.5. We might have a high degree of confidence that our bicycle is outside our apartment in Paris where we left it. That should, if we are coherent in the *or-introduction* inference, give us a high degree of confidence that our bicycle is outside our apartment in Paris or in Timbuktu. But we do not have any confidence that, if our bicycle is not outside our apartment in Paris, then it is in Timbuktu. It is much more reasonable to infer that, if our bike is not there, it is somewhere else in Paris after being stolen. Johnson-Laird and Byrne (2002, p. 650) claimed that people always endorse 1.5, but Gilio and Over (2012) have an analysis of when 1.5 and 1.6 are, and are not, reasonable inferences to make, and Over et al. (2010) have supporting results. Because the question of whether people's responses to the *or-to-if* inferences are coherent depends on how the conditional is interpreted, our investigation of these inferences does more than reveal their coherence in general. It also tells us about the modal interpretation of the conditional. If people's judgments are highly incoherent for one interpretation, and yet highly coherent for another, there is an argument in favor of the interpretation that renders their judgments coherent.

Inferences 2.1 and 2.2 are from a conjunction to the conditional. The first has the conjunction as a single premise, whereas the second has the two conjuncts as separate premises. It is easiest to state what is coherent for the single premise inference to the conditional event. By probability theory, $P(p \text{ and } q) = P(p)P(q|p) \leq P(q|p)$. The formula for the coherence of judgments about the premises and conclusion of the 2.2 inference is more complex because it requires taking into account that the premises can covary to different degrees (Kleiter, 2014). The formula for it is reported in Experiment 2 below. Because of this additional complexity in processing coherence for inference 2.2, we wanted to assess whether people's responses complied with coherence more often for 2.1 than for 2.2.

Inferences 2.3 and 2.4 are forms of *and-elimination*, and we have already stated above how coherence is determined for them. Not conforming to coherence for this inference is to commit the conjunction fallacy. We therefore wanted to test for this case whether our general prediction holds: that people's probability judgments more often conform to coherence when they are given the explicit task of drawing inferences.

EXPERIMENT 1

METHODS

Participants

A total of 1140 participants from English speaking countries completed the online experiment, in exchange for €0.1. From this initial sample we excluded cases that had the same IP address as a previously recorded participant, cases that provided the same response on all trials, and cases that had a reported age below 12 or above 100[1]. The final sample consisted of 871 participants. Their

[1]A reanalysis of the data excluding the 20 participants with a reported age between 12 and 17 led to the same pattern of significant and non-significant results.

mean age was 35 years (range 12–78). They reported different levels of formal educational training, and 87% reported having "good" or "very good" English language skills.

Material and design

Participants were shown a short scenario describing a person, and then presented with a series of statements about the person. The statements either appeared one at a time on the screen, in random order for each participant in the *statements group*, or the statements were presented in pairs as the premises and conclusions of explicit inferences in the *inferences group*. Participants in the statements group were asked to judge how confident they were in each statement, by typing in a percentage between 0% ("no confidence at all") and 100% ("complete confidence"). Participants in the inferences group were asked to judge how confident they were in the premise of the argument, and then how confident they were in the conclusion, given the premise. Participants in the inferences group used the same percentage scale as those in the statements group to provide their answers.

Two scenarios were varied between participants: The Linda scenario (Tversky and Kahneman, 1983), with the standard description of Linda, and a scenario describing a person conforming to a stereotype quite unlike that of Linda. Below is a sample trial in the statements group and in the inferences group, using the Linda scenario:

Statements group:

Now consider the following statement about Linda:

Please indicate how much confidence you would have in this statement. Please give a percentage rating from 0% (no confidence at all) to 100% (complete confidence).

"Linda votes for the Labour Party or the Green Party"

Inferences group:

Now consider the following argument about Linda:

Next to A please indicate how much confidence you would have in the premise of the argument. Next to B please indicate how much confidence you would have in the conclusion, given the premise. Please give a percentage rating from 0% (no confidence at all) to 100% (complete confidence).

A. "Linda votes for the Labour Party or the Green Party"

B. "Therefore, if Linda does not vote for the Labour Party, then she votes for the Green Party"

In the inferences group, participants judged each inference twice with different contents. The allocation of scenario contents to inferences was counterbalanced across participants, leading to eight different booklets, four for each scenario. In the statements group, each participant rated the entire set of contents created for the relevant scenario, leading to two booklets, one for each scenario. In order to compensate for the difference in sample size between groups resulting from the different number of booklets in each group, we placed a weight on the otherwise random procedure for assigning participants to booklets, such that participants were twice as likely to receive any one of the booklets of the statements group than any one of the booklets of the inferences group. This resulted in sample sizes of $n = 305$ and $n = 566$ for the statements and inferences group, respectively.

Procedure

The experiment took place online using the platform CrowdFlower. On the first screen participants viewed the instructions and a sample trial. The next screen showed the scenario within which the statements, or respectively the inferences, were to be assessed. These then followed, presented one at a time on the screen. A further screen asked for demographical information, and a final screen provided debriefing information. The whole procedure took on average 4.24 min for the statements group and 5.23 min for the inferences group.

RESULTS

We measured above chance compliance with coherence using a method introduced by Evans et al. (under review; see also Pfeifer and Kleiter, 2009). First, we computed the difference between the probability assigned to the conclusion and the probability assigned to the premise. We then computed a binary variable to encode whether this difference indicated that the response was coherent or not. Thus, for or-introduction, 1.1–1.2, and the *if-to-or* inferences, 1.3–1.4, this variable took the value 1 when the difference was positive or 0, and took the value 0 otherwise. For the *or-to-if* inferences, 1.5–1.6, the variable took the value 1 when the difference was 0 or negative, and took the value 0 otherwise. This computation was performed separately for each participant and inference. We call this variable *observed coherence*. Next, we computed the probability of a response being coherent by chance, *chance coherence*. On the assumption that a random response can fall equally likely on any point of the probability scale, the probability of complying to coherence by chance corresponds to the width of the coherence interval. This is a simplifying assumption because there is evidence that people's probability estimates might be biased at the boundaries of the interval, in a way that could lead to higher chance rates for extreme cases (c.f. Stewart et al., 2006). However, we considered a uniform distribution of chance rates a sufficiently accurate approximation to allow an assessment of the hypotheses at hand. On this assumption, if a person assigns for instance a probability of 0.6 to the premise of an *or-introduction* inference, then the probability she assigns to the conclusion is coherent if it falls within the interval between 0.6 and 1. Because the width of this interval is 0.4, the chance rate of conforming to coherence is in this case also 0.4. Finally, we subtracted chance coherence from observed coherence, to obtain a measure of the extent to which responses were coherent at levels above those expected by chance, *above chance coherence*.

The ratings of above chance coherence were submitted to a mixed ANOVA with the between subjects factor of *task* (statements, inferences) and the within subjects factor of *inference* (*or-introduction*, *if-to-or*, and *or-to-if*). Throughout the paper, the Greenhouse–Geisser correction of degrees of freedom for lack of sphericity was used when appropriate, and the Bonferroni–Holm correction of p-values for multiple comparisons was used to define the limit of a significant effect, while reporting the original p-values. The results are depicted in **Figure 1**. The overall intercept was significant, $F(1,869) = 885.29$, $p < 0.001$, $\eta_p^2 = 0.505$, indicating that overall probability judgments were coherent at above chance level. There was also a main effect of inference, $F(1.382,1201.390) = 266.28$, $p < 0.001$, $\eta_p^2 = 0.235$: above chance

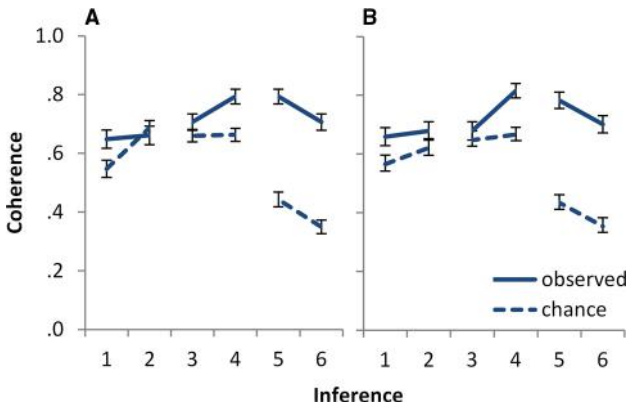

FIGURE 1 | Observed versus chance coherence for the six inferences of Experiment 1, (A) for the statements and (B) for the inferences task. Inferences 1 and 2 are logically equivalent *or-introduction* inferences, with a negation absent in 1 and present in the premise of 2. Inferences 3 and 4 are logically equivalent *if-to-or* inferences. Inference 3 has a negation in the conclusion and inference 4 in the premise. Inferences 5 and 6 are logically equivalent or-to-if inferences. Inference 5 has a negation in the conclusion and inference 6 in the premise. See **Table 1** for the precise logical form of the inferences. Error bars show 95% CI.

coherence differed for the three inferences. No other effects were significant (highest $F = 1.07$, lowest $p = 0.30$). In particular, there was no significant effect of task.

Follow-up analyses of the effect of inference showed that although responses were coherent at above chance level for all three inferences, the degree of above chance coherence was higher for *or-to-if* than for *if-to-or*, $F(1,869) = 270.99$, $p < 0.001$, $\eta_p^2 = 0.238$; and higher for *if-to-or* than for *or-introduction*, $F(1,869) = 16.50$, $p < 0.001$, $\eta_p^2 = 0.019$. Thus, responses to both *or-to-if* and *if-to-or* were consistently above chance. Responses to *or-introduction* were also coherent more often than expected by chance, although somewhat less often than responses to the other two inferences. An inspection of **Figure 1** suggests that the difference between the *if-to-or* and the *or-introduction* inferences was due mainly to the lower coherence for *or-to-if* for inference 1.2 in the statements task. In line with this, a comparison between the two inferences restricted to the inference task showed no difference in above chance coherence between the two, $F(1,565) = 1.85$, $p = 0.17$, $\eta_p^2 = 0.003$.

We conducted a further analysis of the *or-to-if* inference in which we excluded responses that are coherent for both the conditional event and the material conditional interpretation of the conditional: responses that assigned the same probability to the premise and conclusion. We treated as coherent only those responses that are coherent for a conditional event interpretation: responses that assigned a lower probability to the conclusion than to the premise. On a material conditional interpretation, the only coherent response to this inference is to assign the same probability to both the premise and conclusion, and assuming that people interpret the conditional as the material conditional, the mean difference between premise and conclusion probability would be expected to be 0. There might be some scattering of probabilities above and below 0, but no systematic drift in any direction. We would expect there to be no effect of coherence for this analysis. On a conditional event interpretation, responses are coherent when the probability of the conclusion of the *or-to-if* inference is equal to or lower than that of the premise. On this interpretation, we would expect coherence to be lower for this analysis than for the analysis using all the data, because a subset of coherent responses would not be considered. The absence of an effect of coherence would also be compatible with this interpretation, and would then render the analysis uninformative to the question at hand. However, a remaining effect of coherence in the expected direction would constitute specific evidence for a conditional event interpretation and against a material conditional interpretation of the conditional.

An univariate ANOVA on above chance coherence for the *or-to-if* inference, using only the data for which probability judgments differed for premise and conclusion in each individual case, yielded a significant intercept, $F(1,362) = 100.27$, $p < 0.001$, $\eta_p^2 = 0.217$: responses to *or-to-if* were coherent at levels above chance when only considering as coherent those responses that are coherent for the conditional event and incoherent for the material conditional interpretation of the conditional.

Although **Figure 1** shows the results separately for each position of the negation, we did not find any consistent effects regarding this variable. We also did not have any hypotheses about it, but introduced it only as a control variable, to be able to obtain a pattern of results that could be generalized across positions of the negation.

DISCUSSION

We investigated the extent to which people's probability judgments were coherent for the premises and conclusions of inferences 1.1 to 1.6, when these were separate statements, in the statements group, and when they were formed into explicit inferences, in the inference group. We found people's responses to be coherent at levels above chance for the three inferences forms investigated, 1.1–1.2, 1.3–1.4, and 1.5–16, in both the statements group and the inferences group. There was therefore clear evidence that people's probability judgments conform to Bayesian principles, and at the same time there was no evidence that this conformity was improved further in the context of explicit inference.

Responses for the *or-introduction* inferences, 1.1–1.2, were found to be coherent at levels above chance, and to a degree similar to that for the *if-to-or* inferences 1.3–1.4., implying that participants endorse this inference when they are asked for their degrees of belief and not whether, as in a binary experiment, the conclusion necessarily follows given the premise. This finding is in accordance with our prediction that pragmatic factors have less effect on this inference when people are asked for their degrees of belief. Also supporting this conclusion, (Politzer and Baratgin, under review) found, using an ordinal response format for degrees of belief, that responses for *or-introduction* were coherent to a level comparable to five other valid inferences. But they also found coherence rates for the inference to be lower when the premise was certain than when it was uncertain. The limiting case of certainty, which is in effect the only one studied in a binary approach, may give a misleading picture of how

far people conform to Bayesian standards, and this hypothesis will have to be investigated further. One option would be through a comparison of responses with binary and with continuous response format. Although a mapping of the two response scales is not straightforward, larger differences between them would still be informative (see Markovits and Handley, 2005; Singmann and Klauer, 2011, for two ways of carrying out such a comparison).

The analysis of responses to the *or-to-if* inferences, 1.5–1.6, showed that participants' responses would fail to be coherent at levels above chance under a material conditional interpretation of the conditional, *P(if p then q) = P(not-p or q)*, whereas they would be coherent at levels above chance if the natural language conditional is interpreted as the conditional event, *P(if p then q) = P(q|p)*. There does not appear to be any reason why people would be so highly incoherent for these inferences if they had a material conditional understanding of the natural language conditional. But if they have a conditional event understanding, our finding is to be expected, and it provides new support for the conditional event interpretation of the conditional. People's conditional reasoning can be much "improved" from a Bayesian point of view, if their understanding of the conditional is, to begin with, correctly identified by the psychology of reasoning.

Responses to the *if-to-or* inferences were likewise reliably coherent above chance levels, showing that participants respected the difference in the coherence conditions between these and the *or-to-if* inferences.

Experiment 1 investigated inferences between conditionals and disjunctions. Our second experiment addresses inferences between conditionals and conjunctions, and includes the content that is famous for causing the incoherence of the conjunction fallacy.

EXPERIMENT 2

METHOD

Participants

Forty-eight students from the University of Orsay, France, took part in the experiment on a voluntary basis. Their mean age was 20 years (range 18–24). They had different majors, although the majority studied biology or medicine. All participants were French native speakers.

Material and design

The material and design were very similar to those of Experiment 1. However, only the Linda scenario was used, and because the original inferences contained no negations, no negation effects were assessed. Inferences 2.1 and 2.2, *and-to-if* forms, used contents prototypical for the scenario, in order to obtain higher probability estimates for the premises and thus lower probabilities of conforming to coherence just by chance. Inferences 2.3 and 2.4, *and-elimination* inferences, varied the prototypicality of the content for the scenario in the same way as in Tversky and Kahneman's (1983) original work on the conjunction fallacy. To take an example from the explicit inferences group, participants read the standard description of Linda and were then asked to state what confidence they had in "Linda is banker" as a conclusion explicitly inferred from "Linda is a feminist and a banker" as a premise.

Participants were divided into two groups of equal size. The booklets for the statements group contained a continuous list of statements. In the booklets for the inferences group, each inference appeared on a separate page. Four booklets were constructed for each group, which differed only in the order in which the items were presented.

Procedure

Participants were tested in the university library in small groups of up to four participants. They worked at their own pace, and took 10 to 15 min to complete.

RESULTS

Responses to inference 2.1 are incoherent when the probability assigned to the conclusion is lower than that of the premise. For inference 2.2, the computation of coherence is more complicated because it takes into account the minimum and maximum overlap between the two premises. The lower and upper coherence bounds for this inference, when the entailed conditional is interpreted as the conditional event, are as follows:

$$P(conclusion) = P(q|p) \in \left[\max\left\{0, \frac{p+q-1}{p}\right\}, \min\left\{\frac{q}{p}, 1\right\}\right]$$

We only computed coherence for the conditional event interpretation of the conditional. However, the coherence bounds for this interpretation are stronger than those for the material conditional. Therefore, any response that is coherent for the above interpretation is also coherent for the material conditional interpretation. See Politzer (2014) for a proposal of how to obtain the intervals of coherence for a wide range of inferences in an intuitive way using a water tank analogy.

The results are illustrated in **Figure 2**. To assess whether the additional complexity of processing coherence for 2.2 as compared to 2.1 leads to higher levels of above chance coherence for 2.1, we conducted a mixed ANOVA on above chance coherence with the between subjects factor of task (statements, inferences)

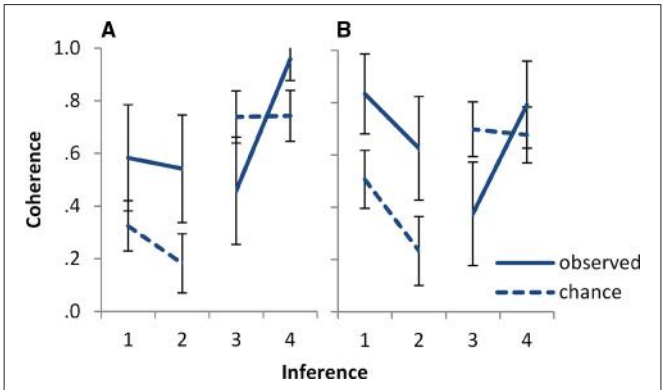

FIGURE 2 | Observed versus chance coherence for the four inferences of Experiment 2, (A) for the statements and (B) for the inferences task. Inferences 1 and 2 are *and-to-if* inferences. The first has the conjunction *p and q* as single premise, the second has *p* and *q* as two separate premises. Inferences 3 and 4 are *and-elimination* inferences. The first has prototypical, and the second counter-prototypical content for the scenario. See **Table 1** for the precise logical form of the inferences. Error bars show 95% CI.

and the within subjects factor of inference (2.1, 2.2). The intercept was significant, $F(1,46) = 36.83$, $p < 0.001$, $\eta_p^2 = 0.445$, indicating that participants' responses were coherent to a degree above that expected by chance. No other effects were significant (largest $F = 1.03$, lowest $p = 0.32$). In particular, there was no significant effect of task or of inference.

To assess whether the conjunction fallacy is reduced in the context of an inference task, a mixed ANOVA on above chance coherence for inferences 2.3 and 2.4 was conducted, with task as a between subjects factor and inference as a within subjects factor. There was a main effect of inference, $F(1,46) = 33.31$, $p < 0.001$, $\eta_p^2 = 0.420$: The two inferences differed in above chance coherence. No other effects were significant (largest $F = 2.12$, smallest $p = 0.15$). In particular, the intercept was not significant, indicating that overall the coherence of participants' judgments for these inferences did not differ from chance; and there was no effect of task. In line with the pattern in **Figure 2**, follow up analyses indicated that coherence was above chance for 2.3, which used prototypical content, $F(1,46) = 10.23$, $p = 0.002$, $\eta_p^2 = 0.182$; and coherence was below chance for 2.4, which used counter-prototypical content, $F(1,46) = 18.37$, $p < 0.001$, $\eta_p^2 = 0.285$.

DISCUSSION

The finding of above chance coherence for the *and-to-if* inferences 2.1 and 2.2 extends the evidence of Experiment 1 to inferences relating conjunctions and conditionals. The absence of an effect of inference implies that at least for the materials used, the additional requirement in 2.2 of integrating two premise probabilities did not reduce the coherence of people's responses. The absence of an effect of task suggests that above chance coherence for this inference was also not affected by the presence of an explicit inference task, similar to the findings from Experiment 1. A further investigation of the extent to which the requirement of integrating premise probabilities affects people's reasoning performance could vary the degree of overlap between premises, as well as the assessment of additional indicators of task difficulty, such as response times.

People's responses to the *and-elimination* inferences 2.3 and 2.4 were coherent at levels above chance, when the materials did not have the content that caused the conjunction fallacy in Tversky and Kahneman (1983). This result is in line with other findings in the probabilistic approach using different methodologies (Pfeifer and Kleiter, 2005; Politzer and Baratgin, under review). However, when the material did have the content known to cause the fallacy, participants were incoherent, just as Tversky and Kahneman would predict for our statements group. Tversky and Kahneman did not predict whether the fallacy would be found when p (or q) was explicitly inferred from p *and* q as a premise. Stating a degree of confidence in the conclusion of such an explicit inference could arguably qualify as what they called a "transparent" problem, to which people should give a coherent answer. Nevertheless, the participants in our inference group were also incoherent by committing the conjunction fallacy, which at least reinforces Tversky and Kahneman's view of it as a deep fallacy that is hard to overcome.

GENERAL DISCUSSION

With the advent of the Bayesian approach in the psychology of reasoning, it has become possible to investigate people's deductive reasoning from uncertain premises, and to assess the extent to which it is coherent. We investigated this topic in two experiments using inferences between conjunctions, disjunctions, and conditionals. We also looked at whether an explicit inference task increases people's coherence, and examined a number of more specific hypotheses for the individual inferences. People's probability judgments were coherent at levels above chance for almost all the inference forms investigated. The one exception was when the materials for the *and-elimination* inference were of the content known to cause the conjunction fallacy. The participants, who read the standard description of Linda, were incoherent in their judgments about "Linda is a feminist and a banker" and "Linda is a banker," even when they inferred the later statement from the former in an explicit inference.

People were generally coherent, complying with the axioms of probability theory, not only in the explicit inference task, but to an equal extent when the task was to evaluate the single statements that formed the inferences one at a time in random order. This absence of an effect of task was not expected. On the one hand, it does provide some support for the descriptive adequacy of the principle of coherence, because it increases the generality of its scope. It stands in accordance with findings on good conformity to Bayesian principles in domains outside of reasoning, where tasks are carried out in a more implicit way, like perception and language comprehension (Fiser et al., 2010; Hsu et al., 2011). And it suggests that addressing the question of what improves Bayesian reasoning should not make us lose sight of the many contexts in which conformity to Bayesian principles is already quite good.

On the other hand, it remains a plausible hypothesis that explicit inference can be an effective use of cognitive resources to improve coherence, to the benefit of reasoning and decision making. The inference forms we considered here, for conjunctions, disjunctions, and their relations to conditionals, may generally be too simple for an effect to be found. Evans et al. (under review) did find that an explicit inference task could increase coherence in a study of MP, MT, AC, and DA. One possibility is that these two-premise conditional inferences require a more complex integration of premise probabilities, and people could be helped to achieve this in explicit inference tasks.

Another possibility is that it was generally easier in the experiment of Evans and et al. (under review) to detect an increase in above chance coherence because the mean probability estimates given to the premises in their experiment were generally higher than in our experiments. Generally, the higher the probability of the premises, the lower the chance rate of coherence and thus the easier it becomes to detect above chance coherence when it is there. This relation holds for MP, MT, AC, and DA, and all the inferences investigated here except for the *or-to-if* inferences, 1.5–1.6, in Experiment 1. For 1.5–1.6, the opposite relation holds: the chance rate of coherence becomes lower, and the probability of detecting above chance coherence larger, the lower the probability assigned to the premise. Because the mean probability ratings for the premises of the inferences in Experiment 1 was relatively low, chance rate coherence was lower for 1.5 and 1.6 than for the

other two inferences, 1.1–1.2 and 1.3–1.4. This explains the higher ratings of above chance coherence for 1.5–1.6 compared to those for *if-to-or* inferences, 1.3–1.4, in spite of comparable rates of observed coherence for both inferences. This also explains why the effect of above chance coherence was relatively small for 1.1–1.2 and 1.3–1.4 in spite of their sizeable rates of observed coherence.

Overall, the dependence of chance rate coherence on the probabilities assigned to the premises is relevant for the interpretation of the presence or absence of incremental effects of coherence, predicted in this case by the presence of an inference task. But it is also relevant to the interpretation of above chance coherence taken by itself, as well as for the interpretation of differences in above chance coherence between inferences. Future experiments on these questions could therefore aim at more adequate control of the premise probabilities, either by letting them be provided by the experimenter, or by conducting a larger pre-test of materials and selecting those with similar probabilities.

As noted above, the high coherence rates described for 1.5–1.6 *or-to-if* inferences, displayed in **Figure 1**, depend on the conditional being interpreted as the conditional event. If the natural language conditional corresponded to the material conditional, with *P(if p then q) = P(not-p or q)* as implied by mental model theory (Byrne and Johnson-Laird, 2009), then the responses to 1.5–1.6 would be incoherent at levels above chance. These results provide strong evidence for the conditional event interpretation of the conditional, and highlight the importance of taking into account people's semantic interpretation of the premises and conclusions for assessing how far they conform to Bayesian principles.

The results from Experiment 2 on the *and-elimination* inferences 2.3 and 2.4 demonstrate that above chance coherence for these forms depends on there being no conflict between the probability of a statement and its contextual prototypicality. It is remarkable how slight variations in these factors can lead to incoherent judgments that resist even explicit inferences. It is a challenge to all accounts of the conjunction fallacy to explain why it persists through apparently "transparent" *and-elimination* inferences. The very reliability of this finding highlights the relevance of investigating further what is driving the conjunction fallacy (see Jarvstad and Hahn, 2011; Oaksford, 2013; Pothos and Busemeyer, 2013; Tentori et al., 2013, for a recent discussion). However, it is more remarkable still that when such conflicts are not present, people give generally coherent probability judgments even in the absence of explicit inference tasks, at least for conjunctions, disjunctions, and their relations to conditionals. This provides further evidence of the descriptive adequacy of Bayesian reasoning principles.

ACKNOWLEDGMENTS

Financial support for this work was provided by a grant from the ANR Chorus 2011 (project BTAFDOC), and by a scholarship from the German Academic Exchange Service (DAAD). We would like to thank Audrey Chansel for adapting the materials for Experiment 2 into French, and collecting the data for this experiment.

REFERENCES

Adams, E. (1998). *A Primer of Probability Logic*. Stanford, CA: CLSI Publications.

Baratgin, J., Over, D. E., and Politzer, G. (2013). Uncertainty and the de Finetti tables. *Think. Reason.* 19, 308–328. doi: 10.1080/13546783.2013.809018

Baratgin, J., Over, D. E., and Politzer, G. (2014). New psychological paradigm for conditionals and general de Finetti tables. *Mind Lang.* 29, 73–84. doi: 10.1111/mila.12042

Bar-Hillel, M., and Neter, E. (1993). How alike is it versus how likely is it: a disjunction fallacy in probability judgments. *J. Pers. Soc. Psychol.* 65, 1119–1131. doi: 10.1037/0022-3514.65.6.1119

Braine, M. D. S., Reiser, B. J., and Rumain, B. (1984). Some empirical justification for a theory of natural propositional reasoning. *Psychol. Learn. Motiv.* 18, 313–337. doi: 10.1016/S0079-7421(08)60365-5

Byrne, R. M. J., and Johnson-Laird, P. N. (2009). 'If' and the problems of conditional reasoning. *Trends Cogn. Sci.* 13, 282–287. doi: 10.1016/j.tics.2009.04.003

Byrne, R. M. J., and Johnson-Laird, P. N. (2010). "Conditionals and possibilities," in *Cognition and Conditionals: Probability and Logic in Human Thought*, eds M. Oaksford and N. Chater (Oxford: Oxford University Press), 55–68. doi: 10.1093/acprof:oso/9780199233298.003.0003

Chater, N., and Oaksford, M. (eds). (2008). *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford: Oxford University Press.

Cruz, N., and Oberauer, K. (2014). Comparing the meanings of "if" and "all." *Mem. Cogn.* 42, 1345–1356. doi: 10.3758/s13421-014-0442-x

de Finetti, B. (1974). *Theory of Probability*. Chichester: John Wiley & Sons.

de Finetti, B. D. (1995). The logic of probability, trans. B. Angell. *Philos. Stud.* 77, 181–190 (Original work published 1936).

Douven, I., and Verbrugge, S. (2010). The Adams family. *Cognition* 117, 302–318. doi: 10.1016/j.cognition.2010.08.015

Edgington, D. (1995). On conditionals. *Mind* 104, 235–329. doi: 10.1093/mind/104.414.235

Elqayam, S., and Over, D. E. (2013). New paradigm psychology of reasoning: an introduction to the special issue edited by Elqayam, Bonnefon, and Over. *Think. Reason.* 19, 249–265. doi: 10.1080/13546783.2013.841591

Espino, O., and Byrne, R. M. J. (2013). The compatibility heuristic in non-categorical hypothetical reasoning: inferences between conditionals and disjunctions. *Cogn. Psychol.* 67, 98–129. doi: 10.1016/j.cogpsych.2013.05.002

Evans, J. St. B. T., Handley, S. J., and Over, D. E. (2003). Conditionals and conditional probability. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 321–335. doi: 10.1037/0278-7393.29.2.321

Evans, J. St. B. T., and Over, D. E. (2004). *If*. Oxford: Oxford University Press.

Evans, J. St. B. T., and Over, D. E. (2013). Reasoning to and from belief: deduction and induction are still distinct. *Think. Reason.* 19, 267–283. doi: 10.1080/13546783.2012.745450

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.* 14, 119–130. doi: 10.1016/j.tics.2010.01.003

Fugard, A. J. B., Pfeifer, N., and Mayerhofer, B. (2011a). Probabilistic theories of reasoning need pragmatics too: modulating relevance in uncertain conditionals. *J. Pragmat.* 43, 2034–2042. doi: 10.1016/j.pragma.2010.12.009

Fugard, A. J. B., Pfeifer, N., Mayerhofer, B., and Kleiter, G. D. (2011b). How people interpret conditionals: shifts towards the conditional event. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 635–648. doi: 10.1037/a0022329

Gilio, A., and Over, D. E. (2012). The psychology of inferring conditionals from disjunctions: a probabilistic study. *J. Math. Psychol.* 56, 118–131. doi: 10.1016/j.jmp.2012.02.006

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Griffiths, T. L., and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychol. Sci.* 17, 767–773. doi: 10.1111/j.1467-9280.2006.01780.x

Hsu, A. S., Chater, N., and Vitányi, P. M. B. (2011). The probabilistic analysis of language acquisition: theoretical, computational, and experimental analysis. *Cognition* 120, 380–390. doi: 10.1016/j.cognition.2011.02.013

Jarvstad, A., and Hahn, U. (2011). Source reliability and the conjunction fallacy. *Cogn. Sci.* 35, 682–711. doi: 10.1111/j.1551-6709.2011.01170.x

Johnson-Laird, P. N., and Byrne, R. M. J. (1991). *Deduction*. Hove: Erlbaum.

Johnson-Laird, P. N., and Byrne, R. M. J. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychol. Rev.* 109, 646–678. doi: 10.1037/0033-295X.109.4.646

Kleiter, G. (2014). "Modeling imprecise degrees of belief by distributions," in *Proceedings of the 8th London Reasoning Workshop*, ed V. Thompson. (London: Birkbeck, University of London).

Markovits, H., and Handley, S. (2005). Is inferential reasoning just probabilistic reasoning in disguise? *Mem. Cogn.* 33, 1315–1323. doi: 10.3758/BF03193231

Oaksford, M. (2013). Quantum probability, intuition, and human rationality. *Behav. Brain Sci.* 36, 203. doi: 10.1017/S0140525X12003081

Oaksford, M., and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Oaksford, M., and Chater, N. (2012). Dual processes, probabilities, and cognitive architecture. *Mind Soc.* 11, 15–26. doi: 10.1007/s11299-011-0096-3

Oaksford, M., and Chater, N. (2013). Dynamic inference in everyday conditional reasoning in the new paradigm. *Think. Reason.* 19, 346–379. doi: 10.1080/13546783.2013.808163

Oaksford, M., Chater, N., and Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 883–899. doi: 10.1037/0278-7393.26.4.883

Oaksford, M., and Hahn, U. (2007). "Induction, deduction and argument strength in human reasoning and argumentation," in *Inductive Reasoning*, eds A. Feeney and E. Heit (Cambridge: Cambridge University Press), 269–301.

Oberauer, K., Geiger, S. M., and Fischer, K. (2011). "Conditionals and disjunctions," in *The Science of Reason: A Festschrift for Jonathan St. B.T. Evans*, eds K. Manktelow, D. E. Over, and S. Elqayam (Hove: Psychology Press), 93–118.

Oberauer, K., Geiger, S. M., Fischer, K., and Weidenfeld, A. (2007). Two meanings of "if"? Individual differences in the interpretation of conditionals. *Q. J. Exp. Psychol.* 60, 790–819. doi: 10.1080/17470210600822449

Oberauer, K., and Wilhelm, O. (2003). The meaning(s) of conditionals: conditional probabilities, mental models, and personal utilities. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 680–693. doi: 10.1037/0278-7393.29.4.680

Over, D. (in press). "The paradigm shift in the psychology of reasoning," in *Human Rationality: Thinking Thanks to Constraints*, eds L. Macchi, M. Bagassi, and R. Viale (Cambridge, MA: The MIT Press).

Over, D. E., Evans, J. St. B. T., and Elqayam, S. (2010). "Conditionals and non-constructive reasoning," in *Cognition and Conditionals: Probability and Logic in Human Thinking*, eds M. Oaksford and N. Chater (Oxford: Oxford University Press), 135–151.

Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., and Sloman, S. A. (2007). The probability of causal conditionals. *Cogn. Psychol.* 54, 62–97. doi: 10.1016/j.cogpsych.2006.05.002

Pfeifer, N. (2013). The new psychology of reasoning: a mental probability logical perspective. *Think. Reason.* 19, 329–345. doi: 10.1080/13546783.2013.838189

Pfeifer, N. (2014). Reasoning about uncertain conditionals. *Stud. Log.* 102, 849–866. doi: 10.1007/s11225-013-9505-4

Pfeifer, N., and Kleiter, G. D. (2005). Coherence and nonmonotonicity in human reasoning. *Synthese* 146, 93–109. doi: 10.1007/s11229-005-9073-x

Pfeifer, N., and Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *J. Appl. Log.* 7, 206–217. doi: 10.1016/j.jal.2007.11.005

Pfeifer, N., and Kleiter, G. D. (2010). "The conditional in mental probability logic," in *Cognition and Conditionals: Probability and Logic in Human Thinking*, eds M. Oaksford and N. Chater (Oxford: Oxford University Press), 153–173.

Pfeifer, N., and Kleiter, G. D. (2011). "Uncertain deductive reasoning," in *The Science of Reason: A Festschrift for Jonathan St B T Evans*, eds K. I. Manktelow, D. E. Over, and S. Elqayam (Hove: Psychology Press), 145–166.

Politzer, G. (2014). *Deductive Reasoning Under Uncertainty: A Water Tank Analogy*. Available at: http://jeannicod.ccsd.cnrs.fr/ijn_00867284 [accessed May 22, 2014].

Politzer, G., Over, D. A., and Baratgin, J. (2010). Betting on conditionals. *Think. Reason.* 16, 172–197. doi: 10.1080/13546783.2010.504581

Pothos, E. M., and Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behav. Brain Sci.* 36, 255–274. doi: 10.1017/S0140525X12001525

Ramsey, F. P. (1994). "Truth and probability," in *Philosophical Papers*, ed. D. H. Mellor (Original work published 1926, Cambridge: Cambridge University Press), 52–94.

Rips, L. (1983). Cognitive processes in propositional thinking. *Psychol. Rev.* 90, 38–71. doi: 10.1037/0033-295X.90.1.38

Rips, L. (1994). *The Psychology of Proof: Deductive Reasoning in Human Thinking*. Cambridge, MA: The MIT Press.

Singmann, H., and Klauer, K. C. (2011). Deductive and inductive conditional inferences: two modes of reasoning. *Think. Reason.* 17, 247–281. doi: 10.1080/13546783.2011.572718

Singmann, H., Klauer, K. C., and Over, D. (2014). New normative standards of conditional reasoning and the dual-source model. *Front. Psychol.* 5:316. doi: 10.3389/fpsyg.2014.00316

Stalnaker, R. C. (1968). "A theory of conditionals," in *Studies in Logical Theory*, ed. N. Rescher (Oxford: Basil Blackwell), 98–112.

Stewart, N., Chater, N., and Brown, G. D. A. (2006). Decision by sampling. *Cogn. Psychol.* 53, 1–26. doi: 10.1016/j.cogpsych.2005.10.003

Tentori, K., Crupi, V., and Russo, S. (2013). On the determinants of the conjunction fallacy: probability vs inductive confirmation. *J. Exp. Psychol. Gen.* 142, 235–255. doi: 10.1037/a0028770

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90, 293–315. doi: 10.1037/0033-295X.90.4.293

Tversky, A., and Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychol. Rev.* 101, 547–567. doi: 10.1037/0033-295X.101.4.547

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 December 2014; paper pending published: 22 January 2015; accepted: 06 February 2015; published online: 25 February 2015.

*Citation: Cruz N, Baratgin J, Oaksford M and Over DE (2015) Bayesian reasoning with ifs and ands and ors. Front. Psychol. **6**:192. doi: 10.3389/fpsyg.2015.00192*

This article was submitted to Cognition, a section of the journal Frontiers in Psychology.

frontiers in Psychology

GENERAL COMMENTARY
published: 27 May 2015
doi: 10.3389/fpsyg.2015.00718

Corrigendum: Bayesian reasoning with ifs and ands and ors

Nicole Cruz[1]*, *Jean Baratgin*[2,3], *Mike Oaksford*[1] and *David E. Over*[4]

[1] Department of Psychological Sciences, Birkbeck, University of London, London, UK, [2] Laboratory CHArt (PARIS), Université Paris 8, Paris, France, [3] Institut Jean Nicod, Paris, France, [4] Department of Psychology, Durham University, Durham, UK

Keywords: uncertain reasoning, deduction, conditionals, coherence, conjunction fallacy

OPEN ACCESS

Edited by:
David R. Mandel,
Toronto Research Centre, Canada

Reviewed by:
Igor Douven,
Paris-Sorbonne University, France

***Correspondence:**
Nicole Cruz,
ncruzd01@mail.bbk.ac.uk

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *13 May 2015*
Accepted: *13 May 2015*
Published: *27 May 2015*

Citation:
Cruz N, Baratgin J, Oaksford M and Over DE (2015) Corrigendum: Bayesian reasoning with ifs and ands and ors. Front. Psychol. 6:718. doi: 10.3389/fpsyg.2015.00718

A corrigendum on

Bayesian reasoning with ifs and ands and ors

by Cruz, N., Baratgin, J., Oaksford, M., and Over, D. E. (2015). Front. Psychol. 6:192. doi: 10.3389/fpsyg.2015.00192

In the article "Bayesian reasoning with ifs and ands and ors," by Nicole Cruz, Jean Baratgin, Mike Oaksford, and David E. Over (*Frontiers in Psychology*, 2015, Vol. 6, Art. 192), on page 6, **Figure 2**, the x-axis in both panels would have to read "1, 2, 4, 3" in order to correctly represent the figure. Rearranging the figure to retain an x-axis labeling in ascending order, the corrected **Figure 2** is displayed as follows.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

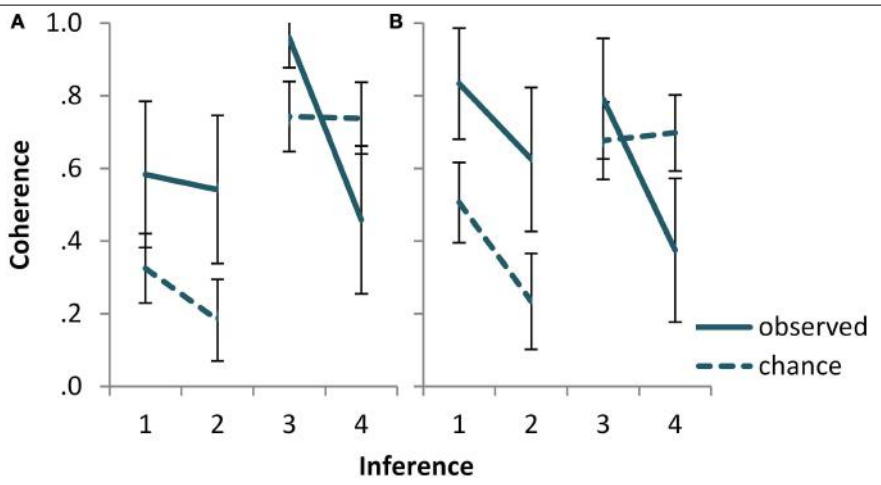

FIGURE | 2 Observed vs. chance coherence for the four inferences of Experiment2, (A) for the statements and (B) for the inferences task. Inferences 1 and 2 are *and-to-if* inferences. The first has the conjunction *p and q* as single premise, the second has *p* and *q* as two separate premises. Inferences 3 and 4 are *and-elimination* inferences. The first has prototypical, and the second counter-prototypical content for the scenario. See Table 1 for the precise logical form of the inferences. Error bars show 95% CI.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 April 2015
doi: 10.3389/fpsyg.2015.00459

Probabilistic alternatives to Bayesianism: the case of explanationism

Igor Douven[1]* *and Jonah N. Schupbach*[2]

[1] Sciences, Normes, Décision, Paris-Sorbonne University, Paris, France, [2] Department of Philosophy, University of Utah, Salt Lake City, UT, USA

There has been a probabilistic turn in contemporary cognitive science. Far and away, most of the work in this vein is Bayesian, at least in name. Coinciding with this development, philosophers have increasingly promoted Bayesianism as the best normative account of how humans ought to reason. In this paper, we make a push for exploring the probabilistic terrain outside of Bayesianism. Non-Bayesian, but still probabilistic, theories provide plausible competitors both to descriptive and normative Bayesian accounts. We argue for this general idea via recent work on explanationist models of updating, which are fundamentally probabilistic but assign a substantial, non-Bayesian role to explanatory considerations.

Keywords: Bayesianism, explanation, updating, inference, probability

OPEN ACCESS

Edited by:
David R. Mandel, Defence Research and Development Canada, Toronto Research Centre, Canada

Reviewed by:
David E. Over, Durham University, UK
Tania Lombrozo, University of California, Berkeley, USA

***Correspondence:**
Igor Douven, Sciences, Normes, Décision, Paris-Sorbonne University, Maison de la Recherche, 28 rue Serpente, 75006 Paris, France
igor.douven@paris-sorbonne.fr

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *14 January 2015*
Paper pending published: *12 February 2015*
Accepted: *30 March 2015*
Published: *27 April 2015*

Citation:
Douven I and Schupbach JN (2015) Probabilistic alternatives to Bayesianism: the case of explanationism. Front. Psychol. 6:459. doi: 10.3389/fpsyg.2015.00459

1. Introduction

There has been a probabilistic turn in the cognitive sciences, a development most prominently marked by the emergence of the "Bayesian paradigm" in the psychology of human learning and reasoning (e.g., Evans and Over, 2004; Griffiths and Tenenbaum, 2006; Tenenbaum et al., 2006; Gopnik and Tenenbaum, 2007; Oaksford and Chater, 2007, 2013; Over, 2009; Baratgin et al., 2013; Elqayam and Evans, 2013) and recent work on the "Bayesian brain" in cognitive neuroscience (e.g., Doya et al., 2006; Friston and Stephan, 2007; Hohwy, 2013). The vast majority of such work is—as in the examples cited above—described by adherents as "Bayesian." In general, probabilistic and Bayesian approaches are so closely associated by cognitive scientists that it rarely is observed that these two approaches may come apart.

There are, nonetheless, various ways in which a theory might be probabilistic without being Bayesian. Most obviously, theories can draw upon probabilities interpreted in non-Bayesian ways (e.g., Gigerenzer and Hoffrage, 1995; Mayo, 1996; Williamson, 2010). But a theory can easily conflict with Bayesianism, even while adopting the standard Bayesian interpretation of probabilities (as measures of agent credences). In this paper, we want to highlight the potential merits of probabilistic, non-Bayesian accounts of this latter sort.

We focus our sights on the question of how humans update their confidences when confronted with new information[1]. Bayesian accounts model such updating strictly in accordance with

Bayes's Rule. Upon learning $A \in \mathcal{A}$ and nothing else between times t_1 and t_2, an agent's credences are to be updated so as to satisfy the equality $\Pr_{t_2}(B) = \Pr_{t_1}(B \mid A)$ for all propositions $B \in \mathcal{A}$, provided $\Pr_{t_1}(A) > 0$.

[1] In this paper, we use "update" in the general sense of belief change. It is worth noting that some authors in the Bayesian camp (e.g., Walliser and Zwirn, 2002; Baratgin and Politzer, 2011) use the term to designate a *particular type* of belief change.

Here, $\mathcal{A}$ is an algebra of propositions over which the probability measures $\Pr_{t_1}$—representing the agent's credences at time t_1—and $\Pr_{t_2}$—representing the agent's credences at later time t_2—are defined, and $\Pr_{t_1}(B \mid A)$ designates the prior (at t_1) conditional probability of B given A.

The Bayesian account thus requires updates to be determined *purely* by an agent's prior conditional (subjective) probabilities. Probabilistic accounts more generally aim to model updating with the help of probability theory. Such accounts may accord with Bayes's Rule, but they need not. A non-Bayesian probabilistic account may, for example, calculate updated credences as a function of prior conditional probabilities plus some other set of factors (probabilistically explicable or not). In the following, we will be especially concerned with "explanationist" models of updating that take explanatory considerations into account in addition to prior conditional probabilities.

There are two crucially distinct ways one can interpret any theory of updating: as providing norms that updates rationally ought to satisfy, or as a descriptive model of how people in fact update. At the same time that cognitive scientists focusing on the descriptive interpretation have increasingly turned to probabilistic models, more and more philosophers have come to regard Bayesianism as providing the norms of both rational action and rational belief (e.g., Maher, 1993; Jeffrey, 2004; Joyce, 2009). Against this seemingly growing consensus on the nature of rationality, the present paper makes a push for exploring the probabilistic terrain outside of Bayesianism and challenges the thought that any deviation from Bayesianism implies a form of irrationality.

A central contention of this paper is that some probabilistic models of updating that conflict with Bayes's Rule constitute strong, plausible competitors to Bayes's Rule, whether the models in question are interpreted descriptively or normatively. We make a case for this claim by focusing on a particular family of non-Bayesian, probabilistic models of updating, namely explanationist models. We argue that explanationist models may be predictively more accurate than Bayesianism (Section 3) without being normatively defective in any way (Section 4). Probabilistic alternatives to Bayesianism accordingly deserve more explicit attention in cognitive science and philosophy than they have thus far received. Before making our case, however, in the next section we offer a general description of explanationism.

2. Explanationism

Deductive inference plays a key role in human reasoning. It is unsurprising, therefore, that this form of inference has been amply studied by psychologists (see, e.g., Evans, 1982; Evans and Over, 1996). Early on, psychologists commonly regarded deductive logic as providing standards of rational reasoning. But psychologists eventually came to realize that not all reasoning proceeds by deductive inference, and that the issue of rationality can arise also for forms of reasoning that are of a non-deductive nature. Having seen hundreds of white swans without ever having seen a swan of a different color, we may infer that all swans are white. While—as we now know—this inference would be to a false conclusion, it is not obviously irrational, and certainly more rational than if we inferred the same conclusion on the basis of having seen a mere handful of white swans, or after already having encountered a black swan. Indeed, many of our beliefs are seemingly held on the basis of this type of "inductive inference," as it is now commonly called, and many of those beliefs would appear to be *rationally* held on that basis. So, it is again no surprise that there is a vast amount of work on this type of inference to be found in the psychological literature (see, e.g., Rips, 2001; Heit and Feeney, 2005; Heit, 2007; Heit and Rotello, 2010).

What *is* surprising is the almost complete neglect by psychologists of a form of inference that is neither deductive nor inductive but that does seem to play a key role—for better or worse—in human thinking. The form of inference we mean has been labeled "abductive inference" (or "abduction") by the great American pragmatist philosopher Charles Sanders Peirce. (See the supplement on Peirce of Douven, 2011 for references). Abduction and induction distinguish themselves from deduction by being *ampliative*: unlike deductively valid inferences, cogent abductive and inductive arguments do not guarantee the truth of a conclusion on the basis of the truth of the premises. Abduction then distinguishes itself from induction by giving pride of place to explanatory considerations, in that it makes the believability of a hypothesis partly a matter of how well the hypothesis explains the available evidence.

To illustrate, consider the following famous anecdote about the invasion of the Thames by the Dutch fleet in 1667—also known as "the Raid on the Medway"—and Sir Isaac Newton, who was a Fellow at Trinity College, Cambridge, at the time:

> Their guns were heard as far as Cambridg, and the cause was well-known; but the event was only cognizable to Sir Isaac's sagacity, who boldly pronounc'd that they had beaten us. The news soon confirm'd it, and the curious would not be easy whilst Sir Isaac satisfy'd them of the mode of his intelligence, which was this; by carefully attending to the sound, he found it grew louder and louder, consequently came nearer; from whence he rightly infer'd that the Dutch were victors. [William Stukeley, *Memoirs of Sir Isaac Newton's Life*, quoted in Westfall (1980 p. 194)]

The "mode of intelligence" referred to here, which according to Westfall's (1980, p. 194) struck the other Fellows in Cambridge with awe, is most plausibly thought of as involving abductive reasoning. It is exceedingly difficult to think of a reasonable set of premises—reasonable from Newton's perspective at the time—from which the conclusion that the Dutch had won follows deductively. Nor did the Dutch—or any other nation that possessed a sizable fleet in the second half of the seventeenth century—invade England frequently enough for Newton's reasoning to be naturally construed as inductive. Rather, it seems that what led Newton to his conclusion is that a Dutch victory was the *best explanation* for his evidence: there are various potential explanations of why the sound of the canon fire grew louder and louder that do not involve a Dutch victory. For instance, the British fleet might have defeated the Dutch, but then that victory might have been followed by a mutiny in which the British marines turned against their own headquarters. However, this and other alternative potential explanations are topped, in terms

of explanatory goodness, by the hypothesis that the Dutch fleet had beaten the British.

Abduction has been identified as playing a central role in scientific reasoning by various historians and philosophers of science (e.g., McMullin, 1984, 1992; Lipton, 1993, 2004; Achinstein, 2001). McMullin (1992) even refers to abduction as "the inference that makes science." This is not to say that abduction has no place outside of science. Various authors have argued for its prominence in everyday contexts as well, for instance, that abductive reasoning is routinely and automatically invoked when we rely on the words of others (Harman, 1965; Adler, 1994; Fricker, 1994) and even in interpreting the words of others (e.g., Bach and Harnish, 1979, p. 92; Hobbs, 2004). In philosophy, abduction has been relied on in defenses of the position of scientific realism, according to which science progressively succeeds in providing better and better representations of reality (Boyd, 1984; Psillos, 1999), as well as in defenses of various metaphysical theses (e.g., Shalkowski, 2010).

A more modern name for abduction is "Inference to the Best Explanation" (IBE), and most statements of abduction to be found in the literature are rather straightforward unpackings of that name. In Musgrave's (1988, p. 239) formulation, for instance, abduction is the principle according to which "[i]t is reasonable to accept a satisfactory explanation of any fact, which is the best available explanation of that fact, as true," and Psillos (2004, p. 83) tells us that "IBE authorizes the acceptance of a hypothesis H, on the basis that it is the best explanation of the evidence." Such formulations raise questions of their own. What makes one explanation better than others? When is an explanation satisfactory? And, ought we really to accept the best explanation of the evidence even if it explains the evidence very poorly? Moreover, one wonders what the relationship between abduction and Bayesianism might be, given that abduction is apparently stated in terms of the categorical notion of acceptance, and does not refer to probabilities or credences.

In recent years, researchers have become interested in a version of abduction that is probabilistic in nature and even has Bayes's Rule as a limiting case (Douven, 2013; Douven and Wenmackers, in press). Where $\{H_i\}_{i\leqslant n}$ is a set of self-consistent, mutually exclusive, and jointly exhaustive hypotheses, this version of abduction models human learning as an act of updating one's degrees of belief on new evidence in accordance with

Probabilistic abduction. Upon learning $E \in \mathcal{A}$ and nothing else between times t_1 and t_2, an agent's credences are to be updated so as to satisfy the equality

$$\Pr_{t_2}(H_i) \;=\; \frac{\Pr_{t_1}(H_i)\Pr_{t_1}(E \mid H_i) + \mathcal{E}(H_i, E)}{\sum_{j=1}^{n}\big(\Pr_{t_1}(H_j)\Pr_{t_1}(E \mid H_j) + \mathcal{E}(H_j, E)\big)},$$

with $\mathcal{E}$ assigning a bonus to the hypothesis that explains the evidence best, and nothing to the other hypotheses, and supposing $\Pr_{t_1}(E) > 0$.

It is easy to verify that probabilistic abduction concurs with Bayes's Rule if $\mathcal{E}$ is set to be the constant function 0, meaning that no bonus points for explanatory bestness are ever attributed. It is not much more difficult to verify that probabilistic abduction concurs with Bayes's Rule *only if* no bonus points are assigned (Douven and Wenmackers, in press).

Naturally, as stated here, probabilistic abduction is really only a schema as long as $\mathcal{E}$ has not been specified. For present purposes, this matter can be left to the side. In fact, for this paper, the rule only serves to show that there are versions of abduction that are direct contenders to Bayes's Rule. But one can think of many more probabilistic update rules that explicate the broad idea that explanatory considerations have confirmation-theoretic import—the central idea underlying abduction. Rather than advocating any particular such rule, we now proceed to argue that the Bayesian model of updating—whether construed descriptively or normatively—may plausibly be improved in various ways by taking into account explanatory considerations, leaving the details of how exactly to account for such considerations for another occasion.

3. Explanationism vs. Bayesianism: Descriptive Adequacy

Contrary to what the growing popularity of Bayesianism among psychologists might lead one to expect, studies regularly find that people update in ways inconsistent with the Bayesian model; see, for instance, Phillips and Edwards (1966), Robinson and Hastie (1985), and Zhao et al. (2012)[2]. What is more, there is evidence suggesting that explanatory considerations do have an impact on people's beliefs; see, for instance, Koehler (1991); Pennington and Hastie (1992); Josephson and Josephson (1994); Thagard (2000); Lombrozo (2006, 2007, 2012); Lombrozo and Carey (2006); Douven and Verbrugge (2010); Bonawitz and Lombrozo (2012); Legare and Lombrozo (2014), and Lombrozo and Gwynne (2014).

The typical reaction to such findings is to look on departures from Bayesian reasoning as a complication or problem, and subsequently to hunt for explanations for why people are ostensibly straying from the proper rational norms. A far less explored option is to question whether Bayes's Rule (and with it Bayesianism) describes the appropriate normative standard for updating. We ask the normative question in the next section. In this section, we explore whether probabilistic models that take into account explanatory considerations might do better at describing people's updating behavior than Bayes's Rule.

The non-Bayesian, probabilistic models that we examine are related to research reported in Douven and Schupbach (in press), which in turn built on research reported in Schupbach (2011). The focus of the latter paper was on probabilistic measures of explanatory goodness or "power," which aim to formalize the degree to which a potential explanation H accounts for evidence E. For example, according to a very simple proposal, H explains E to a degree equal to $\Pr(E \mid H) - \Pr(E)$. Other—prima facie more promising—measures that have been discussed in the

[2]This is not to deny that there is also evidence in support of the descriptive adequacy of Bayesianism. See in particular Griffiths and Tenenbaum (2006); Tenenbaum et al. (2006); Gopnik and Tenenbaum (2007), and Oaksford and Chater (2007).

philosophy of science literature include Popper's (1959) measure,

$$\frac{\Pr(E \mid H) - \Pr(E)}{\Pr(E \mid H) + \Pr(E)},$$

Good's (1960) measure,

$$\ln\left(\frac{\Pr(E \mid H)}{\Pr(E)}\right),$$

and Schupbach and Sprenger's (2011) measure,

$$\frac{\Pr(H \mid E) - \Pr(H \mid \neg E)}{\Pr(H \mid E) + \Pr(H \mid \neg E)}.$$

It is to be noticed that, while all three measures have 0 as the "neutral point," they are not all on the same scale. In particular, Popper's and Schupbach and Sprenger's measures have range $[-1, 1]$ while Good's measure has range $(-\infty, \infty)$. However, Schupbach (2011) also considers functional rescalings of Good's measure obtained via this schema:

$$L_\alpha(x) = \begin{cases} 1 - e^{-x^2/2\alpha^2} & \text{if } x \geqslant 0; \\ -1 + e^{-x^2/2\alpha^2} & \text{if } x < 0, \end{cases}$$

which do all have range $[-1, 1]$. Below, we use "L_a" to refer to the rescaling of Good's measure obtained in this way with $\alpha = a$.

Schupbach (2011) sought to answer the question of how well these and some other measures of explanatory goodness capture people's judgments of explanatory goodness. To that end, an experiment was conducted in which 26 participants were individually interviewed. In the interviews, the participants were shown two urns containing 40 balls each, with one urn ("urn A") containing 30 black balls and 10 white ones, and the other urn ("urn B") containing 15 black balls and 25 white ones. Each interview started by informing the participant about the contents of the urn and giving him or her a visual representation of these contents—which remained in sight during the whole interview. The experimenter then tossed a fair coin and decided, based on the outcome, whether urn A or urn B would be chosen. The participant knew that an urn was chosen in this way, but was not informed about which urn had been selected. Instead, the experimenter drew 10 balls from the selected urn, without replacement, and lined up the drawn balls in front of the participant. After each draw, participants were asked: (i) to judge the explanatory goodness, in light of the draws so far, of the hypothesis that urn A had been selected (H_A); (ii) to do the same for the hypothesis that urn B had been selected (H_B); and (iii) to assess how likely it was in the participant's judgment that urn A had been selected, given the outcomes at that point. The participant had to answer the questions about explanatory goodness by making a mark on a continuous scale with five labels at equal distances, the leftmost label reading that the hypothesis at issue was an extremely poor explanation of the evidence so far, the rightmost reading that the hypothesis was an extremely good explanation, and the labels in between reading that the hypothesis was a poor/neither poor nor good/good explanation, in the obvious order.

The data obtained in this experiment allowed Schupbach to calculate, for each participant and for each of the measures that he considered, the explanatory power of H_A and H_B after each draw the participant had witnessed, where either objective probabilities or credences could be used for the calculations. The results of these calculations were compared with the actual judgments of explanatory goodness that the participant had given after each draw. The results somewhat favored Schupbach and Sprenger's (2011) measure over its competitors. In general, however, Popper's measure, various rescalings of Good's measure, and Schupbach and Sprenger's measure all performed well in predicting participant judgments concerning explanatory power—regardless of whether explanatory power was calculated on the basis of objective probabilities or on the basis of credences.

In Douven and Schupbach (in press), the data gathered in Schupbach's experiment were re-analyzed for a very different purpose. Whereas Schupbach used credences as well as objective probabilities to calculate values of explanatory goodness according to the above measures, which were then compared with participants' judgments of explanatory goodness, Douven and Schupbach were instead interested in the role that such judgments play in updating credences. Put differently, where Schupbach took judgments of explanatory goodness to be the response variable and either credences or objective probabilities as the input for one of the measures of explanatory goodness, the output of which then served as the predictor variable, Douven and Schupbach took credences as the response variable and objective probabilities and judgments of explanatory goodness as possible predictors. In doing so, they hoped to shed light on the question of the role of explanatory considerations in updating, in particular, of whether taking into account such considerations, possibly in conjunction with objective probabilities, leads to better predictions of people's updates—as should be the case, according to the descriptive reading of explanationism.

To be more precise, Douven and Schupbach (in press) first collected the credences of all participants into one variable (call this variable "S"), the objective conditional probabilities that those credences should have matched for the updates on the draws to obey Bayes's Rule into a second variable (call this "O"), the judgments of explanatory goodness of H_A into a third variable ("A"), and the judgments of explanatory goodness of H_B into a fourth ("B"). They then fitted a number of linear regression models, with S as response variable and with all or some of O, A, and B as predictor variables. The most interesting comparison was between the Bayesian model (called "MO" in the paper), which had only O as a predictor variable, and the full, explanationist model ("MOAB"), which had O, A, and B as predictor variables. In this comparison, as in the general comparison between all models that had been fitted, the explanationist model clearly came out on top. The difference in AIC value between MO and MOAB was over 120 in favor of the latter. Also, MOAB had an R^2 value of 0.90, while MO had an R^2 value of 0.83. A likelihood ratio test also favored MOAB over MO: $\chi^2_{(2)} = 124.87$, $p < 0.0001$.

In short, the explanationist model MOAB was much more accurate in predicting people's updates than the Bayesian model MO, strongly suggesting that, at least in certain contexts, agents's explanatory judgments play a significant role in influencing how

they update. Note that, by accepting this conclusion, one is not leaving the probabilistic paradigm: conditional probabilities figure as a highly significant predictor in MOAB as well. The conclusion *is* strongly non-Bayesian, however, insofar as MOAB identifies explanatory judgments as significant predictors, too, in conflict with what ought to hold if people were strict Bayesian updaters.

The previous research showed that, in a context in which one is trying to predict people's updated credences, if next to objective probabilities one has access to people's explanatory judgments, one is well-advised also to take the latter into account. In reality, however, we rarely know people's explanatory judgments. Does explanationism suggest anything helpful in contexts in which only objective probabilities are available? It may well do so. Provided we have all the probabilistic information at hand that is required as input for the measures of explanatory power stated above, we can use the output of those measures in combination with objective probabilities and try to predict someone's updates on that combined basis. Given that Schupbach (2011) found a number of the measures of explanatory power to capture well people's judgments of explanatory power, and given that Douven and Schupbach (in press) found people's judgments of explanatory power to co-determine significantly their subjective probabilities, there is reason to believe that objective probabilistic information alone allows one to improve upon Bayesian models, which ignore explanatory considerations altogether.

In Douven and Schupbach (in press), only judgments of explanatory goodness were taken into account; no degrees of explanatory goodness determined by any measure of explanatory power were considered. To see whether such degrees of explanatory goodness (derived from the objective probabilistic information available) help make more accurate predictions about people's updates, we had another look at the data from Schupbach (2011) and fitted a series of linear models similar to MOAB, but now with participants's judgments of explanatory goodness replaced with calculated degrees of explanatory goodness. Specifically, we constructed linear models with S as response variable and O, degrees of explanatory goodness of H_{A}, and degrees of explanatory goodness of H_{B} as predictors. Values of the last two predictors were determined in five distinct ways: using Popper's measure, using three separate rescalings of Good's measure ($L_{0.5}$, L_1, L_2), and using Schupbach and Sprenger's measure. In the following, variable "$\mathrm{Y_X}$" represents degrees of explanatory goodness for hypothesis H_{Y} ($\mathrm{Y} \in \{\mathrm{A}, \mathrm{B}\}$) calculated using measure $\mathrm{X} \in \{\mathrm{P}, \mathrm{G1}, \mathrm{G2}, \mathrm{G3}, \mathrm{SS}\}$, where "P" stands for Popper's measure, "G1" for $L_{0.5}$, which is the first rescaled version of Good's measure, and so on. Similarly, "MXYZ" names the model with predictors X, Y, and Z.

Table 1 gives some important statistics for comparing the models, where we have also included MO from Douven and Schupbach (in press). Because MO is nested within each of the other models, it could be compared with them by means of likelihood tests. The χ^2 column in **Table 1** gives the outcomes of these tests, which were all in favor of the richer model. Given that the χ^2 values obtained in the tests were all significant, this is a first indication that any of the explanationist models provides a better fit with the data than the Bayesian model. Naturally, the better fit might be due precisely to the fact that the explanationist models include more predictors than MO. For that reason, it is worth looking also at the AIC metric, which weighs model fit and model complexity against each other and penalizes for additional parameters. Burnham and Anderson (2002, p. 70) argue that a difference in AIC value greater than 10 indicates that the model with the higher value enjoys basically no empirical support. It is plain to see that MO has a higher AIC value than any of the other models, where the difference is always greater than 10 except in the case of the last model.

Furthermore, we see that it makes a large difference which measure is used to calculate degrees of explanatory goodness. In particular, the model which includes next to O also A_{G3} and B_{G3} as predictors—so degrees of explanatory goodness obtained via L_2—does best: it has the lowest AIC value of all models, the difference each time being greater than 10, and it has the highest R^2 value (although in this respect all models are close to each other). This is confirmed by applying closeness tests for non-nested models to pairs of models consisting of $\mathrm{MOA}_{G3}\mathrm{B}_{G3}$ and one of the other explanationist models. Using Vuong's (1989) model, $\mathrm{MOA}_{G3}\mathrm{B}_{G3}$ is significantly preferred over any of the other explanationist models (in each case, $p < 0.01$), except for $\mathrm{MOA}_{G1}\mathrm{B}_{G1}$; in a comparison of $\mathrm{MOA}_{G3}\mathrm{B}_{G3}$ with $\mathrm{MOA}_{G1}\mathrm{B}_{G1}$, Vuong's test has no preference for either model. On the other hand, using Clarke's (2007) test, we find that $\mathrm{MOA}_{G3}\mathrm{B}_{G3}$ is preferred over *all* other explanationist models (in each case, $p < 0.0001$). **Table 2** gives the regression results for $\mathrm{MOA}_{G3}\mathrm{B}_{G3}$. That O, A_{G3}, and B_{G3} are all highly significant buttresses Douven and Schupbach's (in press) suggestion that when people receive new evidence, they change their credences not only on the basis

TABLE 1 | Comparison of seven regression models.

	k	LL	AIC	ΔAIC	χ^2	R^2
MO	3	202.39	−398.77	48.06		0.83
$\mathrm{MOA}_P\mathrm{B}_P$	5	222.64	−435.27	11.55	40.50***	0.85
$\mathrm{MOA}_{G1}\mathrm{B}_{G1}$	5	216.72	−423.43	23.40	28.66***	0.85
$\mathrm{MOA}_{G2}\mathrm{B}_{G2}$	5	211.27	−412.53	34.29	17.76**	0.84
$\mathrm{MOA}_{G3}\mathrm{B}_{G3}$	5	228.41	−446.83	0.00	52.06***	0.86
$\mathrm{MOA}_{SS}\mathrm{B}_{SS}$	5	208.27	−406.53	40.30	11.76*	0.84

*k is the number of parameters and LL the log-likelihood of each model. AIC is the Akaike Information Criterion, an index for model selection that takes model fit (i.e., log-likelihood) and model complexity (i.e., number of parameters) into account. ΔAIC is the AIC value minus the smallest AIC value. Models with smaller indices provide a more parsimonious (i.e., better) description of the data. R^2 is the squared correlation between the fitted and observed values. *p < 0.05, **p < 0.01, ***p < 0.001.*

TABLE 2 | Regression results for the best explanationist model $\mathrm{MOA}_{G3}\mathrm{B}_{G3}$.

Variable	*B*	*SE B*	*β*	*t*	*p*
Intercept	0.33	0.02		14.90	< 0.0001
O	0.40	0.04	0.56	9.72	< 0.0001
A_{G3}	0.24	0.03	0.30	7.48	< 0.0001
B_{G3}	−0.13	0.03	−0.15	−3.67	0.0002

of objective probabilistic considerations, but also on the basis of explanatory considerations. (At least, it supports that claim in the light of Schupbach's (2011) findings, which indicate a close match between subjective judgments of explanatory goodness and degrees of explanatoriness as calculated by any of the measures at issue.)

Finally, it is worthwhile comparing $MOA_{G3}B_{G3}$ (the best model with degrees of explanatory goodness determined via L_2) with MOAB [the best model from Douven and Schupbach (in press) incorporating recorded judgments of explanatory goodness]. As previously remarked, the R^2 value of MOAB equals 0.90. Its AIC value equals −519.64. So, on both counts, MOAB does better. MOAB is also preferred over $MOA_{G3}B_{G3}$ according to Vuong's test ($p < 0.001$) as well as according to Clarke's test ($p < 0.0001$). This implies that, *if* judgments of explanatory goodness are at hand, then one does best to take them into account in predicting people's updates. As noted, however, very often one will not have a choice, inasmuch as judgments of explanatory goodness are typically unavailable.

In fact, if judgments of explanatory goodness are available, one can even consider constructing a model that includes *both* variables encoding those judgments *and* variables encoding degrees of explanatory goodness, for instance, based on L_2. Doing this for the present case, we find that in a model with all of O, A, B, A_{G3}, and B_{G3}, as predictors, B_{G3} is no longer significant. However, the model with the remaining variables as predictors does significantly better than MOAB in a likelihood ratio test: $\chi^2_{(1)} = 9.12$, $p = 0.003$. Also, the expanded model has a lower AIC value: −526.76. The R^2 value is the same (0.90) for both models.

Summing up, we have found evidence that, at least in some contexts, explanationism is descriptively superior to Bayesianism: by taking explanatory considerations into account, next to conditional probabilities, we arrive at more accurate predictions of people's updates than we would on the basis of the objective conditional probabilities alone. Naturally, the kind of context we considered is rather special, and more work is needed to see how far the results generalize. Nonetheless, our results weigh against the generality of the increasingly popular hypothesis that people tend to update by means of Bayes's Rule.

4. Explanationism vs. Bayesianism: Normative Adequacy

Here is a natural response to the findings of the previous section: "Surely people's updates do indeed break with Bayes's Rule. But this is unsurprising. Bayes's Rule is best interpreted as a norm of proper or rational updating in the light of new evidence. It is an idealization that actual agents can at best hope to approximate, to the extent that they are reasoning as they should. Even if experimental evidence calls descriptive Bayesianism into question then, it does nothing to invalidate Bayesianism as an ideal, normative theory." In this section, we challenge this idea, summarizing recent work that compares Bayes's Rule with explanationist models of updating in order to clarify their respective roles in a full normative theory of rational updating.

Consider the so-called dynamic Dutch Book argument, which has convinced many philosophers that Bayes's Rule is the only rational update rule[3]. This argument has concomitantly done much to discredit explanationism as a normative account. The argument proceeds by describing a collection of bets, some of which are offered to a non-Bayesian updater before that person's update on new information and some of which are offered to him or her after that event. The claim is that, whatever the specifics of the update rule used by the person (other than that it deviates from Bayes's Rule), the pay-offs of the bets can be so chosen that all of them will appear fair in the eyes of the updater at the moment they are offered, yet jointly they ensure a negative net pay-off (such a collection of bets is called "a dynamic Dutch book"). This betokens irrationality on the updater's part—it is claimed—given that the updater could have seen the loss coming. Conversely, it is argued that had the person updated via Bayes's Rule, he or she could not have deemed all bets in the dynamic Dutch book to be fair.

There are at least three reasons for being dissatisfied with this argument. First, Douven (1999) points out that, in the dynamic Dutch book argument, what makes the non-Bayesian updater vulnerable to a dynamic Dutch book is not the use of a non-Bayesian update rule *per se*, but rather the combination of that rule and certain decision-theoretic principles, notably ones for determining the fairness of bets. As argued in the same paper, update rules must be assessed not in isolation, but as parts of *packages* of rules, which include decision-theoretic rules and possibly further update rules. Making use of a decision-theoretic principle proposed in Maher (1992), Douven demonstrates the existence of packages of rules that include a non-Bayesian update rule but that nevertheless do not leave one susceptible to dynamic Dutch books.

Second, even if non-Bayesian updating did make one vulnerable to dynamic Dutch books, it would not follow that such updating is necessarily irrational. For the possibility has *not* been ruled out that non-Bayesian updating has advantages that outweigh any risk of suffering financial losses at the hands of a Dutch bookie. It has recently been shown, in the context of a coin-tossing model in which it is unknown whether the coin is biased and if so what bias it has, that by updating via probabilistic abduction, one is on average faster—virtually always *much* faster—in attributing a high probability (explicated as a probability above 0.09, for instance) to the true bias hypotheses than if one updates via Bayes's Rule (Douven, 2013). Various philosophers have argued that high probability is a necessary condition for rational assertion and action: to be warranted in asserting or acting upon a proposition, the proposition must be highly probable. What this means is that a non-Bayesian scientist may get in a position to assert (including publish) the outcomes of his or her research more quickly than a Bayesian scientist who is working on the same theoretical problems. Or a non-Bayesian stock trader may be sooner warranted in making a profitable buy or sell than the Bayesians on the floor are, simply because he or she

[3]The dynamic Dutch book argument was first published by Teller (1973), who attributed it to David Lewis. Lewis's handout containing the argument was later published (Lewis, 1999).

is quicker in assigning a high probability to the hypothesis that a given firm is going to do very well (or very poorly). Hence, for all Bayesians have shown, even if non-Bayesian updater's expose themselves to Dutch bookies, the financial losses they thereby risk incurring may be more than compensated for in other ways—*inter alia*, non-Bayesians's credences may converge toward the truth more quickly than those of their Bayesian competitors.

Third, even many Bayesians have become dissatisfied with the dynamic Dutch book argument. Above, it was said that the argument heavily depends also on what decision-theoretic principles are assumed. However, such principles would seem out of place in debates about *epistemic* rationality, which concern what it is rational to *believe*, or how to rationally change one's *beliefs* or *credences*, and not how it is rational to *act*. When we talk about rational action (e.g., the rationality of buying a bet), the notion of rationality at play is that of *practical* or *prudential* rationality. Even if Bayesian updating were the rational thing to do, practically speaking, it would not follow that it is the rational thing to do, epistemically speaking.

Motivated by this concern, Bayesians have sought to give an altogether different type of defense of their update rule. The alternative approach starts from the idea that update rules, like epistemic principles in general, are to be judged in light of their conduciveness to our epistemic goal(s), and that it is epistemically rational to adopt the update rule that is most likely to help us achieve our epistemic goal(s). The defense adopts inaccuracy minimization as the preeminent epistemic goal; update rules are accordingly epistemically defensible to the extent that they allow us to minimize the inaccuracy of our credences—where inaccuracy is spelled out in terms of some standard scoring rule(s). And according to Bayesians, it is their favored update rule that does best in this regard[4].

It has recently been noted, however, that the goal of inaccuracy minimization, as it is used in the previous defense, is multiply ambiguous (Douven, 2013). That one ought to minimize the inaccuracy of one's credences can be interpreted as meaning that every update ought to minimize *expected* inaccuracy, but also as meaning that every update ought to minimize *actual* inaccuracy, or again differently, that every update ought to contribute to the long-term project of coming to have a minimally inaccurate representation of the world. And if understood in the third sense, there is the further question of whether we should aim to have minimally inaccurate degrees of belief in the long run, irrespectively of how long the run may be, or whether we should aim at some reasonable trade-off between speed of convergence and precision (see Douven, 2010).

What has effectively been shown is that Bayes's Rule minimizes inaccuracy in the first sense. However, no argument has been provided for holding that minimizing inaccuracy in that sense trumps minimizing inaccuracy in one of the other senses. So, in light of results showing that, given these other interpretations of our epistemic goal, certain versions of abduction outperform Bayes's Rule in achieving that goal (Douven, 2013; Douven and Wenmackers, in press), the inaccuracy minimization defense fails.

The upshot is that there is currently no good reason to hold that Bayesianism describes the unequivocally superior normative theory of updating. Both arguments that implore us to believe otherwise—the dynamic Dutch book argument and the inaccuracy minimization argument—fail in this regard. Bayes's Rule may be the uniquely best at enabling us to achieve one particular epistemic goal (minimizing expected inaccuracy in the long run). But there are other epistemic goals that we might have, which also involve the minimization of inaccuracy and which seem equally legitimate. Relative to some of these, abduction proves to be more conducive than Bayes's Rule. Results reported in Douven (2013) suggest that the precise epistemic goal(s) we should seek to satisfy is a matter that depends on context. That would mean that in some contexts Bayes's Rule is the preferred choice while in others it is abduction. But that is enough reason to reject the idea that abduction is an aberrant update rule, generally inferior to Bayes's Rule.

5. Conclusion

Nothing that we have said here calls into question the value of the probabilistic turn in recent cognitive science. We do, however, take issue with the narrowness of the focus of work in this vein. While we think that there is much fruit to be gleaned from modeling (actual and ideal) credences using probabilities, doing so does not necessitate using a Bayesian account. We have strived here to exemplify a promising way to expand fruitful research being pursued in cognitive science and philosophy today: namely, by exploring the probabilistic terrain outside of Bayesianism.

Doing so, we found strong support for explanationism, both as a descriptive and normative theory. At least in certain contexts, people do seem to base their updates partly on explanatory considerations; and at least with respect to certain plausible epistemic ends, that is what they ought to do. The present Research Topic (in which this article has been placed) centers around the question of how to improve Bayesian reasoning. This question could be taken to presuppose that Bayesianism is the one apt model of uncertain reasoning, and that all departures from Bayesianism are in need of improvement, repair, or explaining-away. In the above, we have challenged these presuppositions. Our findings suggest that when people update their credences partly on the basis of explanatory considerations and thereby flout Bayesian standards of reasoning, that can be because doing so puts them in a better position to achieve their epistemic goals. So, at least in some contexts, we can improve *upon* Bayesianism by taking into account the explanatory merits or demerits of the objects of our credences. To put the message in different terms, instead of asking how to motivate people to reason more in accordance with Bayesian standards, we should ask whether making people more Bayesian is a good idea to begin with.

We suspect that the answer to this question will depend sensitively on context and on the specific epistemic goals that are most salient for an epistemic agent. More research is thus needed to explore when exactly people are non-Bayesians and when exactly they should be. Specifically, do people tend to rely on some

[4]See Rosenkrantz (1992) for an influential early attempt along these lines; it also contains a detailed exposition of scoring rules.

version of abduction mostly in those contexts in which it is best for them to do so, and similarly for Bayes's Rule? Bradley (2005, p. 362) argues that Bayes's Rule "should not be thought of as a universal and mechanical rule of updating, but as a technique to be applied in the right circumstances, as a tool in what Jeffrey terms the 'art of judgment'." Indeed, a key element in the art of judgment may be the ability to judge when to rely on Bayes's Rule and when to rely on abduction or other rules. In addition to this, it may comprise the art of judging explanatory goodness, which also means: not perceiving explanations where there are none. As with every art, one would expect some people to be better at this than others. (As an anonymous referee rightly noted, conspiracy theorists are inclined to see explanations everywhere, and abductive reasoning is likely to hamper rather than help such people to achieve their epistemic goals.)

While the above is not a call to abandon Bayes's Rule across the board—in some contexts, it may be exactly the right rule to follow—our present findings do go straight against Bayesianism as philosophers commonly understand that position, namely, as the position that any deviance from Bayesian updating betokens irrationality. It is to be emphasized, however, that there is no apparent incompatibility between our findings and much of the work in psychology that commonly goes under the banner of Bayesianism. There is nothing in the writings of Chater, Evans, Oaksford, Over, or most of the other researchers commonly associated with the Bayesian paradigm in psychology that obviously commits them either to Bayes's Rule as a universal normative principle or to the hypothesis that, as a matter of fact, people generally do obey the rule[5]. Oaksford and Chater (2013, p. 374) are quite explicit in this regard when they end their discussion of belief change in the context of the new Bayesian paradigm in psychology with the remark that "it is unclear what are the rational probabilistic constraints on dynamic inference." We hope to have shed some new light on this matter by showing that, at least in some contexts, we do well to heed explanatory considerations, both as epistemic agents and as researchers trying to predict the cognitive behavior of others. More generally, we hope to inspire further research on the descriptive and normative merits of probabilistic, but non-Bayesian accounts of human reasoning.

Acknowledgments

We are greatly indebted to Tania Lombrozo and David Over for valuable comments on a previous version of this paper.

References

Achinstein, P. (2001). *The Book of Evidence*. Oxford: Oxford University Press.

Adler, J. (1994). Testimony, trust, knowing. *J. Philos.* 91, 264–275. doi: 10.2307/2940754

Bach, K., and Harnish, R. (1979). *Linguistic Communication and Speech Acts*. Cambridge, MA: MIT Press.

Baratgin, J., Over, D. E., and Politzer, G. (2013). Uncertainty and the de finetti tables. *Think. Reason.* 19, 308–328. doi: 10.1080/13546783.2013.809018

Baratgin, J., and Politzer, G. (2011). Updating: a psychologically basic situation of probability revision. *Think. Reason.* 16, 253–287. doi: 10.1080/13546783.2010.519564

Bonawitz, E. B., and Lombrozo, T. (2012). Occam's rattle: children's use of simplicity and probability to constrain inference. *Dev. Psychol.* 48, 1156–1164. doi: 10.1037/a0026471

Boyd, R. (1984). "The current status of scientific realism," in *Scientific Realism*, ed J. Leplin (Berkeley, CA: University of California Press), 41–82.

Bradley, R. (2005). Radical probabilism and Bayesian conditioning. *Philos. Sci.* 72, 342–364. doi: 10.1086/432427

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multi-model Inference: A Practical Information-Theoretic Approach*. Berlin: Springer.

Clarke, K. (2007). A simple distribution-free test for nonnested hypotheses. *Polit. Anal.* 15, 347–363. doi: 10.1093/pan/mpm004

Douven, I. (1999). Inference to the best explanation made coherent. *Philos. Sci.* 66, S424–S435. doi: 10.1086/392743

Douven, I. (2010). Simulating peer disagreements. *Stud. Hist. Philos. Sci.* 41, 148–157. doi: 10.1016/j.shpsa.2010.03.010

Douven, I. (2011). "Abduction," in *Stanford Encyclopedia of Philosophy*, ed E. Zalta (Spring 2011). Available online at: http://plato.stanford.edu/entries/abduction/

Douven, I. (2013). Inference to the best explanation, Dutch books, and inaccuracy minimisation. *Philos. Q.* 69, 428–444. doi: 10.1111/1467-9213.12032

Douven, I., and Schupbach, J. N. (in press). The role of explanatory considerations in updating. *Cognition*.

Douven, I., and Verbrugge, S. (2010). The Adams family. *Cognition* 117, 302–318. doi: 10.1016/j.cognition.2010.08.015

Douven, I., and Wenmackers, S. (in press). Inference to the best explanation versus Bayes' rule in a social setting. *Br. J. Philos. Sci.*

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. N. (2006). *Bayesian Brain*. Cambridge, MA: MIT Press.

Elqayam, S., and Evans, J. St. B. T. (2013). Rationality in the new paradigm: strict versus soft Bayesian approaches. *Think. Reason.* 19, 453–470. doi: 10.1080/13546783.2013.834268

Evans, J. St. B. T. (1982). *The Psychology of Deductive Reasoning*. London: Routledge.

Evans, J. St. B. T., and Over, D. E. (1996). *Rationality and Reasoning*. Hove: Psychology Press.

Evans, J. St. B. T., and Over, D. E. (2004). *If*. Oxford: Oxford University Press.

Fricker, E. (1994). "Against gullibility," in *Knowing from Words*, eds B. K. Matilal and A. Chakrabarti (Dordrecht: Kluwer), 125–161.

Friston, K. J., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese* 159, 417–458. doi: 10.1007/s11229-007-9237-y

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiment. *J. R. Stat. Soc.* B22, 319–331.

Gopnik, A., and Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. *Dev. Sci.* 10, 281–287. doi: 10.1111/j.1467-7687.2007.00584.x

Griffiths, T. L., and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychol. Sci.* 17, 767–773. doi: 10.1111/j.1467-9280.2006.01780.x

Harman, G. (1965). The inference to the best explanation. *Philos. Rev.* 74, 88–95. doi: 10.2307/2183532

Heit, E. (2007). "What is induction and why study it?," in *Inductive Reasoning*, eds A. Feeney and E. Heit (Cambridge: Cambridge University Press), 1–24.

Heit, E., and Feeney, A. (2005). Relations between premise similarity and inductive strength. *Psychon. Bull. Rev.* 12, 340–344. doi: 10.3758/BF03196382

Heit, E., and Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 805–812. doi: 10.1037/a0018784

[5]While Bayes's Rule has a very central place in the work of Griffiths, Tenenbaum, and their collaborators (see, e.g., Griffiths and Tenenbaum, 2006; Tenenbaum et al., 2006), even these authors do not commit to the claim that Bayes's Rule is the only rational update rule, or the rule that people everywhere and always use to accommodate new information.

Hobbs, J. R. (2004). "Abduction in natural language understanding," in *The Handbook of Pragmatics*, eds L. Horn and G. Ward (Oxford: Blackwell), 724–741.
Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.
Jeffrey, R. (2004). *Subjective Probability: The Real Thing*. Cambridge: Cambridge University Press.
Josephson, J. R., and Josephson, S. G. (eds.). (1994). *Abductive Inference*. Cambridge: Cambridge University Press.
Joyce, J. (2009). "Accuracy and coherence: prospects for an alethic epistemology of partial belief," in *Degrees of Belief*, eds F. Huber and C. Shmidt-Petri (Dordrecht: Springer), 263–300.
Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychol. Bull.* 110, 499–519. doi: 10.1037/0033-2909.110.3.499
Legare, C. H., and Lombrozo, T. (2014). Selective effects of explanation on learning in early childhood. *J. Exp. Child Psychol.* 126, 198–212. doi: 10.1016/j.jecp.2014.03.001
Lewis, D. (1999). "Why conditionalize?," in *Papers on Metaphysics and Epistemology*, ed D. Lewis (Cambridge: Cambridge University Press), 403–407.
Lipton, P. (1993). Is the best good enough? *Proc. Aristotelian Soc.* 93, 89–104.
Lipton, P. (2004). *Inference to the Best Explanation, 2nd Edn.* London: Routledge.
Lombrozo, T. (2006). The structure and function of explanations. *Trends Cogn. Sci.* 10, 464–470. doi: 10.1016/j.tics.2006.08.004
Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cogn. Psychol.* 55, 232–257. doi: 10.1016/j.cogpsych.2006.09.006
Lombrozo, T. (2012). "Explanation and abductive inference," in *Oxford Handbook of Think. Reason*, eds K. J. Holyoak and R. G. Morrison (Oxford: Oxford University Press), 260–276.
Lombrozo, T., and Carey, S. (2006). Functional explanation and the function of explanation. *Cognition* 99, 167–204. doi: 10.1016/j.cognition.2004.12.009
Lombrozo, T., and Gwynne, N. Z. (2014). Explanation and inference: mechanical and functional explanations guide property generalization. *Front. Neurosci.* 8:700. doi: 10.3389/fnhum.2014.00700
Maher, P. (1992). Diachronic rationality. *Philos. Sci.* 59, 120–141.
Maher, P. (1993) *Betting on Theories*. Cambridge: Cambridge University Press.
Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago, IL: University of Chicago Press.
McMullin, E. (1984). "A case for scientific realism," in *Scientific Realism*, ed J. Leplin (Berkeley, CA: University of California Press), 8–40.
McMullin, E. (1992). *The Inference that Makes Science*. Milwaukee, WI: Marquette University Press.
Musgrave, A. (1988). "The ultimate argument for scientific realism," in *Relativism and Realism in Science*, ed R. Nola (Dordrecht: Kluwer), 229–252.
Oaksford, M., and Chater, N. (2007). *Bayesian Rationality*. Oxford: Oxford University Press.
Oaksford, M., and Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Think. Reason.* 19, 346–379. doi: 10.1080/13546783.2013.808163
Over, D. E. (2009). New paradigm psychology of reasoning. *Think. Reason.* 15, 431–438. doi: 10.1080/13546780903266188
Pennington, N., and Hastie, R. (1992). Explaining the evidence: tests of the story-model for juror decision making. *J. Pers. Soc. Psychol.* 62, 189–206. doi: 10.1037/0022-3514.62.2.189
Phillips, L. D., and Edwards, W. (1966). Conservatism in a simple probability inference task. *J. Exp. Psychol.* 72, 346–354. doi: 10.1037/h0023653
Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson. doi: 10.1037/0096-1523.11.4.443
Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. London: Routledge.
Psillos, S. (2004). "Inference to the best explanation and bayesianism," in *Induction and Deduction in the Sciences*, ed F. Stadler (Dordrecht: Kluwer), 83–91.
Rips, L. J. (2001). Two kinds of reasoning. *Psychol. Sci.* 12, 129–134. doi: 10.1111/1467-9280.00322
Robinson, L. B., and Hastie, R. (1985). Revision of beliefs when a hypothesis is eliminated from consideration. *J. Exp. Psychol. Hum. Percept. Perform.* 11, 443–456.
Rosenkrantz, R. D. (1992). The justification of induction. *Philos. Sci.* 59, 527–539. doi: 10.1086/289693
Schupbach, J. N. (2011). Comparing probabilistic measures of explanatory power. *Philos. Sci.* 78, 813–829. doi: 10.1086/662278
Schupbach, J. N., and Sprenger, J. (2011). The logic of explanatory power. *Philos. Sci.* 78, 105–127. doi: 10.1086/658111
Shalkowski, S. (2010). "IBE, GMR, and metaphysical projects," in *Modality: Metaphysics, Logic, and Epistemology*, eds B. Hale and A. Hoffmann (Oxford: Oxford University Press), 167–187.
Thagard, P. (2000). *How Scientists Explain Disease*. Princeton, NJ: Princeton University Press.
Teller, P. (1973). Conditionalization and observation. *Synthese* 26, 218–258. doi: 10.1007/BF00873264
Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends Cogn. Sci.* 10, 304–318. doi: 10.1016/j.tics.2006.05.009
Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333. doi: 10.2307/1912557
Walliser, B., and Zwirn, D. (2002). Can Bayes' rule be justified by cognitive rationality principles? *Theory Decis.* 53, 95–135. doi: 10.1023/A:1021227106744
Westfall, R. S. (1980). *Never at Rest*. Cambridge: Cambridge University Press.
Williamson, J. (2010). *In Defence of Objective Bayesianism*. Oxford: Oxford University Press.
Zhao, J., Crupi, V., Tentori, K., Fitelson, B., and Osherson, D. (2012). Updating: learning versus supposing. *Cognition* 124, 373–378. doi: 10.1016/j.cognition.2012.05.001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

REVIEW
published: 31 March 2015
doi: 10.3389/fpsyg.2015.00340

Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why

Gary L. Brase[1]* *and W. Trey Hill*[2]

[1] Department of Psychological Sciences, Kansas State University, Manhattan, KS, USA, [2] Department of Psychology, Fort Hays State University, Hays, KS, USA

Bayesian reasoning, defined here as the updating of a posterior probability following new information, has historically been problematic for humans. Classic psychology experiments have tested human Bayesian reasoning through the use of word problems and have evaluated each participant's performance against the normatively correct answer provided by Bayes' theorem. The standard finding is of generally poor performance. Over the past two decades, though, progress has been made on how to improve Bayesian reasoning. Most notably, research has demonstrated that the use of frequencies in a natural sampling framework—as opposed to single-event probabilities—can improve participants' Bayesian estimates. Furthermore, pictorial aids and certain individual difference factors also can play significant roles in Bayesian reasoning success. The mechanics of how to build tasks which show these improvements is not under much debate. The explanations for *why* naturally sampled frequencies and pictures help Bayesian reasoning remain hotly contested, however, with many researchers falling into ingrained "camps" organized around two dominant theoretical perspectives. The present paper evaluates the merits of these theoretical perspectives, including the weight of empirical evidence, theoretical coherence, and predictive power. By these criteria, the ecological rationality approach is clearly better than the heuristics and biases view. Progress in the study of Bayesian reasoning will depend on continued research that honestly, vigorously, and consistently engages across these different theoretical accounts rather than staying "siloed" within one particular perspective. The process of science requires an understanding of competing points of view, with the ultimate goal being integration.

Keywords: Bayesian reasoning, frequencies, probabilities, ecological rationality, heuristics and biases, pictorial aids, numeracy

OPEN ACCESS

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
Laura Felicia Martignon, Ludwigsburg University of Education, Germany
Wim De Neys, Centre National de la Recherche Scientifique, France
Rocio Garcia-Retamero, University of Granada, Spain

***Correspondence:**
Gary L. Brase, Department of Psychological Sciences, Kansas State University, 492 Bluemont Hall, Manhattan, KS 66506, USA
gbrase@ksu.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *06 January 2015*
Accepted: *10 March 2015*
Published: *31 March 2015*

Citation:
Brase GL and Hill WT (2015) Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. Front. Psychol. 6:340. doi: 10.3389/fpsyg.2015.00340

Introduction

Imagine, for one moment, the following scene: A !Kung woman begins her day by foraging for berries in the Kalahari Desert. Wandering from patch to patch, she searches for substantial

portions of subsistence. Foraging is not always fruitful; it does not always yield food, and sometimes it does not yield enough food to justify the calories expended during the act of foraging. Foragers must decipher patterns from the environment in order to be successful and efficient. For example, the !Kung woman may have success 90% of the time she travels to the east canyon, but only when she forages in springtime. During the summer months, the east canyon may be barren of food. At some level of cognition, the woman must coarsely analyze the data from her travels in order to determine the odds of finding food in the east canyon, given the fact that it is springtime or summer. From a psychological perspective, we may wonder what is happening at the cognitive, or algorithmic, level in the woman's mind. How is she storing the information, and how is she arriving at seemingly appropriate solutions to this particular problem of calculating a posterior probability of finding food given certain environmental cues? Although the surface of this paper provides guidance for ways to improve Bayesian reasoning, it also delves into the deeper questions of how and why the mind is designed to solve certain problems with specific inputs.

The General Case of Bayesian Reasoning

The technical name for what the !Kung woman is doing in the above story is *Bayesian reasoning*. Although Bayesian reasoning sometimes has a narrow mathematical definition (i.e., the use of Bayes theorem, specifically), for the purposes of psychological research the more relevant definition is the general process of using new information (e.g., season of the year) to calculate the revised likelihood that an event of a known prior base rate will occur (e.g., successfully finding food). Humans have, historically, needed to perform quick computational estimates of such probabilities in order to navigate various aspects of ancestral environments (Cosmides and Tooby, 1996). Therefore, it seems scientifically unproductive to insist on the narrow definition (in that an explicit Bayes theorem is only a few centuries old) in describing human judgments and decision making. It is important therefore to distinguish between a narrow and rigid usage of "applying Bayes' theorem" in defining Bayesian reasoning, as compared to a more general usage of Bayesian reasoning as a process of adaptively updating prior probabilities with new information (by whatever means) to reach a new, or posterior, probability. This more general definition of Bayesian reasoning, which is the sensible one to take from the perspective of a cognitive psychologist, is to evaluate behaviors as the potential product of cognitive mechanisms acting "as if" they were Bayesian. Specifically, this general definition of Bayesian reasoning can be used to classify behaviors based on the observable evidence that the individual organism in question used new evidence to update its estimate that an event would occur. Often, this is ultimately tested through some measurable behavior (e.g., a decision to act in accordance with this new evidence's implications for the posterior probability of an event).

Bayesian Reasoning as a Serious, Real World Problem

Traditional research on people's abilities to engage in Bayesian reasoning uses the following protocol: a person is presented with a description of a situation in which Bayesian reasoning is relevant, the necessary numerical information for Bayesian calculations, and then a request that the participant calculate the posterior probability (expressed in terms of the relevant situation). For example, one such task (adapted from Chapman and Liu, 2009) is as follows:

> The serum test screens pregnant women for babies with Down's syndrome. The test is a very good one, but not perfect. Roughly 5% of babies have Down's syndrome. If a baby has Down's syndrome, there is a 80% chance that the result will be positive. If the baby is unaffected, there is still a 20% chance that the result will still be positive. A pregnant woman has been tested and the result is positive. What is the chance that her baby actually has Down's syndrome?

Undergraduates, medical students, and even physicians do quite poorly on this type of Bayesian reasoning task (e.g., Casscells et al., 1978; Gigerenzer et al., 2007), including when it is in a medical testing context such as the above example. Such failures of Bayesian reasoning suggest potentially tragic consequences for medical decision making, as well as any other real world topics that involve similar calculations.

Interestingly, evaluations of how and why people do poorly in Bayesian reasoning has changed over the years. In the early days of research on Bayesian reasoning, the dominant view by researchers was that humans were approximating Bayes' theorem, but erred in being far too conservative in their estimates (e.g., Edwards, 1982). That is, people did not utilize the new information as much as they should; relying too much on the base rate information. Later work, however, shifted to the idea that the dominant error was in the opposite direction: that people generally erred in relying too much on the new information and neglecting the base rate, either partially or entirely (e.g., Kahneman and Tversky, 1972; Tversky and Kahneman, 1974, 1982). This later approach is one of the better known positions within what is known as the *heuristics and biases* paradigm, within which base rate neglect was considered so strong and pervasive that at one point it was asserted: "In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all" (Kahneman and Tversky, 1972, p. 450).

Improving Bayesian Reasoning

Nevertheless, research continued on human Bayesian reasoning and how to improve it. Beginning in the 1990s, progress began to occur, followed quickly by theoretical debates. There continue to be disagreements to this day, but there now clearly are certain procedures which do in fact improve human Bayesian reasoning. These include: using a natural sampling structure, using frequencies, and using pictures. Each of these procedures also raise theoretical issues about what cognitive processes underlying

the improvement in human reasoning, and this paper will look at each of these in turn. We will also look at the role of individual differences in aptitude and motivation within the context of Bayesian reasoning before concluding with an overall assessment.

Natural Sampling and Frequencies in Bayesian Reasoning

A seminal paper in terms of improving Bayesian reasoning and the current issues revolving around those improvements is Gigerenzer and Hoffrage (1995). This paper described a structure for presenting information in such a way that it greatly helped people reach correct Bayesian conclusions. This structure is one of whole-number frequencies in a natural sampling framework. (This original paper used the unfortunately ambiguous label of "frequency format" for this structure, which has led to some confusion; see Gigerenzer and Hoffrage, 1999, 2007; Lewis and Keren, 1999; Mellers and McGraw, 1999; Vranas, 2000; Gigerenzer, 2001.) There are thus two aspect of this structure: (a) the use of frequencies as a numerical format, and (b) the use of a particular structure, called natural sampling, for the relationships between the numbers. The rationale for both of these aspects is similar: they map onto the type of information which the human mind generally encounters in the natural environment, both currently and over evolutionary history. For this reason, the Gigerenzer and Hoffrage position is often described as the *ecological rationality* approach.

It can be challenging to dissociate natural sampling from frequencies. When considering the occurrence of objects or events in the real world, that experience tends to strongly imply frequency counts as the format in which that information would be encoded. The actual format of natural sampling, however, is actually the online categorization of that information into groups, including groups which can be subsets of one another. **Figure 1** shows the previously given Bayesian reasoning task information (about a Down's syndrome serum test) as naturally sampled frequencies. In this case we imagine (or recall) 100 experiences with this test, and five of those experiences were with a baby who had Down's syndrome (i.e., 5% base rate). Those five experiences can be further categorized by when the test came out positive (4 times; 4 out of 5 is 80%), and the 95 cases of babies without Down's syndrome can be similarly categorized by the test results (19 false positive results; 19 out of 95 is 20%). This nested categorization structure creates numbers in the lower-most row for which the base-rates (from the initial categorization groups) are automatically taken into account already. This, in turn, makes the calculations for Bayesian reasoning less computationally difficult. (Specifically, the probabilistic version of Bayes theorem is $p(H|D) = p(H)p(D|H)/p(H)p(D|H) + p(\sim H)p(D|\sim H)$, with D = new data and H = the hypothesis, whereas with natural sampling this equation can be simplified to $p(H|D) = d\&h/d\&h + d\&\sim h$, with d&h = frequency of data and the hypothesis and d&∼h = frequency of data and the null hypothesis. Also note that changing the natural frequency numbers to standardized formats, such as percentages, destroys the nested categorizations, and thus the computational simplification, of natural sampling.) Thus, whereas it is pretty easy to create numerical frequencies which are not in a natural sampling framework, it is difficult to construct a natural sampling framework without reference to frequencies.

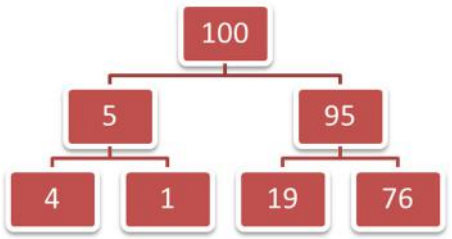

FIGURE 1 | An illustration of a natural sampling framework: the total population (100) is categorized into groups (5/95) and those groups are categorized into parallel sub-groups below that.

The consequences of confusions about how natural sampling and numerical frequencies are related to each other has led to a number of claimed novel discoveries, which are observed from the other side as "re-inventions." One example of this is that the principles of natural sampling have been co-opted as something new and different. These situations require some clarification, which hopefully can be done in a relatively concise manner.

Subsequent to the description and application of a natural sampling structure in the original Gigerenzer and Hoffrage (1995) paper (which explicitly drew on the work by Kleiter (1994) in developing the natural sampling idea), the basic structure of natural sampling has been re-invented at least four times in the literature. Each time, the new incarnation is described at a level of abstraction which allows one to consider the structure independent of frequencies (or any other numerical format), but the natural sampling structure is unmistakable:

(a) Johnson-Laird et al. (1999) reintroduced the basic relevant principle of natural sampling as their "subset principle," implying that ecological rationality researchers somehow missed this property: "The real burden of the findings of Gigerenzer and Hoffrage, (1995, p. 81) is that the mere use of frequencies does not constitute what they call a 'natural sample.' Whatever its provenance, as they hint, a natural sample is one in which the subset relations can be used to infer the posterior probability, and so reasoners do not have to use Bayes' theorem." Note also the confusion in this passage between the narrow definition of Bayesian reasoning as using Bayes' theorem and the more general, psychologically relevant definition of Bayesian reasoning we clarified earlier in this paper. Girotto and Gonzalez (2001) continue from this point in their use of the "subset principle," which is simply an abstraction of the natural sampling structure;

(b) Evans et al. (2000) proposed a process that involves "cueing of a set inclusion mental model," rather than a natural sampling structure;

(c) Macchi (1995) and Macchi and Mosconi (1998) created the label of "partitive formulation" to describe the natural sampling structure; and

(d) Sloman et al. (2003) use the term "nested-set relations" rather than natural sampling, following Tversky and Kahneman (1983).

As this last re-invention noted, Tversky and Kahneman (1983) did discover that using frequencies sometimes improved performance (e.g., in their work on the conjunction fallacy), but they did not actually elaborate this observation into a theory; they only speculated that frequencies somehow helped people represent class inclusion.

Dissociating the natural sampling framework, claiming that it is something else, and then looking at the effects of numerical frequencies by themselves (without natural sampling or with malformed natural sampling) has allowed for all sorts of methodological and conceptual shenanigans. It is not interesting, either methodologically or theoretically, that making Bayesian reasoning tasks harder (by adding steps, using wordings which confuse people, switching numerical formats within the same problem) can decrease performance (see, Brase, 2002, 2008, 2009a,b, 2014 for further elaboration). Indeed, it is generally difficult to make strong theoretical claims based on people failing to accomplish a task, as there are usually many different possible reasons for failure.

In addition to multiple attempts to co-opt the concept of natural sampling there has been a notable attempt to co-opt the numerical format of frequencies, claiming that the facilitative effect of using frequencies is not actually about the frequencies themselves. Girotto and Gonzalez (2001) asserted that people actually can be good at Bayesian reasoning when given only probabilistic information. The probabilities used in this research, however, are of a peculiar type stated in whole number terms. For example:

> Mary is tested now [for a disease]. Out of the entire 10 chances, Mary has ___ chances of showing the symptom [of the disease]; among these chances, ___ chances will be associated with the disease. (p. 274)

How many times was Mary tested? Once or ten times? If tested once, there is one "chance" for a result; if tested 10 times (or even if 10 hypothetical times are envisioned), then this is an example of frequency information. It seems odd to say that subjects are truly reasoning about unique events and that they are not using frequencies, when the probabilities are stated as *de facto* frequencies (i.e., 3 out of 10). Although Girotto and Gonzalez (2001) claim that "chances" refer to the probability of a single-event, it can just as easily be argued that this format yields better reasoning because it manages – in the view of the research participants—to tap into a form of natural frequency representation. This alternative interpretation was immediately pointed out (Brase, 2002; Hoffrage et al., 2002), but advocates of the heuristics and biases approach were not swayed (Girotto and Gonzalez, 2002).

In order to adjudicate this issue, Brase (2008) gave participants Bayesian reasoning tasks based on those used by Girotto and Gonzalez (2001). Some of these problems used the natural sampling-like chances wording. Other versions of this problem used either percentages (not a natural sampling format) or used a (non-chances) frequency wording that was in a natural sampling format. After solving these problems, the participants were asked how they had thought about the information and reached their answer to the problem. First of all, contrary to the results of Girotto and Gonzalez (2001), it was found that frequencies in a natural sampling structure actually led to superior performance over "chances" in a natural sampling structure. (The effect size of this result is actually similar to the Girotto and Gonzalez (2001) results, which were statistically underpowered due to small sample sizes.) More notably, though, the *participants who interpreted the ambiguous "chances" as referring to frequencies performed better than those who interpreted the same information as probabilities.* This result cuts through any issues about the computational differences between natural sampling frameworks versus normalized information, because the presented information is exactly the same in these conditions and requires identical computations; only the participants' understanding of that information is different.

Using Pictures to Aid Bayesian Reasoning

Generally speaking, pictures help Bayesian reasoning. Like the research on frequencies and natural sampling, however, there is disagreement on how and why they help. The ecological rationality account (Cosmides and Tooby, 1996; Brase et al., 1998) considers pictorial representations as helping because they help to tap into the frequency-tracking cognitive mechanisms of a mind designed by the ecology experienced over evolutionary history. That is, people have been tracking, storing, and using information about the frequencies of objects, locations, and events for many generations. Visual representations of objects, events, and locations should therefore be closer to that type of information with which the mind is designed to work. An alternative heuristics and biases account is that pictures help to make the structure of Bayesian reasoning problems easier to understand. This account of pictures helping because it enables people to "see the problem more clearly" is often tied to the co-opted and abstracted idea of natural sampling; the pictures help make the subset structure, the set-inclusion model, or the nested-set relations more apparent (e.g., Sloman et al., 2003; Yamagishi, 2003). Indeed, there are parallels here in the comparison of these two perspectives: the ecological rationality account proposes a more narrowly specified (and evolutionary based) account, whereas the heuristics and biases account favors a less specific (non-evolutionary) account.

Subsequent research (Brase, 2009a, 2014) has taken advantage of the fact that ambiguous numerical formats can be interpreted as either frequencies or as probabilities. By using the "chances" wording for the actual text and therefore holding the numerical information as a constant, while varying the type of pictorial representation, this research has been able to compare different types of pictorial aids against a neutral task backdrop. Brase (2009a) found that, compared to control conditions of no picture at all, Venn circles (which should facilitate the perception of subset relationship) did not help nearly as much as pictures of icon arrays (which should facilitate frequency interpretations of the information). Furthermore, a picture with intermediate properties – a Venn circle with dots embedded within it – led

to intermediate performance between solid Venn circles and icon arrays. Subsequent research by Sirota et al. (2014b) took an interesting intermediate theoretical position, claiming that the heuristics and biases account predicted *no facilitation* of Bayesian reasoning from using pictures (contra Sloman et al., 2003 and Yamagishi, 2003). Their null findings of several different types of pictures failing to improve Bayesian reasoning are used to challenge the ecological rationality account, which they agree does predict an improvement with the use of pictures. A nearly concurrent publication replicated and extended the specific effects of Brase (2009a), however, casting doubt on the significance of the Sirota et al. (2014b) null findings. Brase (2014) found that roulette wheel diagrams (like those used in Yamagishi, 2003) led to performance similar to that of Venn diagrams, and that both realistic and abstract icon shapes significantly improved performance. Interpretation of the ambiguous numerical information as frequencies also improved Bayesian reasoning performance in all these conditions (replicating the findings of Brase, 2008), separate from the effects of the different picture types.

Individual Differences in Bayesian Reasoning

There have been various claims that certain individual differences may moderate the often-observed frequency effect in Bayesian reasoning. Peters et al. (2006) demonstrated that numerical literacy (or numeracy)—an applicable understanding of probability, risk, and basic mathematics—moderated many classic judgment and decision making results, showing proof of concept that not all judgment and decision making tasks may be viewed the same by every individual. Specifically, Peters et al. (2006) showed that low numerates may benefit the most from number formats designed to aid comprehension of the information. The explanation proposed for these results can be summarized as a "fluency hypothesis": that more numerically fluent people (higher in numerical literacy) are influenced less by the use of different numerical formats because they are quite capable of mentally converting formats themselves. In doing so, these highly numerate people utilize the numerical format best suited for the present task. Less numerically fluent people, on the other hand, are prone to work only with the numerical information as presented to them. This leaves them more at the mercy of whatever helpful or harmful format is given to them. Although Peters et al. (2006) did not assess Bayesian reasoning specifically, Chapman and Liu (2009) later brought the issue of numerical literacy to the topic of frequency effects in Bayesian reasoning tasks.

The story takes an interesting turn at this point, because although Peters et al. (2006) showed low numerates benefited most from a number format change to frequencies, Chapman and Liu (2009) showed instead that high numerates differentially benefited from natural frequency formatted Bayesian reasoning problems. Specifically they found that this frequency effect was only observed in highly numerate individuals, resulting in a statistically significant numeracy x number format interaction. Chapman and Liu (2009) pointed out that some other research is consistent with these results. In particular, Bramwell et al. (2006) provided different groups of participants with Bayesian reasoning problems framed as a test for a birth defect. The participants were either obstetricians, pregnant women and their spouses, or midwives. The effect of presentation format was assessed with a between-subjects manipulation, with some participants receiving naturally sampled frequencies and others receiving a single event probability format. Although the frequency effect was observed in their study, a closer examination showed that this effect was limited to obstetricians, whereas the midwives, pregnant women, and their spouses all showed equally poor Bayesian reasoning performance regardless of number format.

To the extent that obstetricians have somewhat higher numerical literacy, which is a plausible assumption, the Bramwell et al. (2006) results would be consistent with those of Chapman and Liu (2009). Both of these results, however, are inconsistent with the findings and the fluency hypothesis of Peters et al. (2006). Chapman and Liu (2009) proposed something akin to a "threshold" hypothesis regarding the interaction effect they found. This threshold hypothesis proposes that a certain level of numerical literacy is required for difficult problems (such as Bayesian reasoning tasks) before helpful formats (e.g., naturally sampled frequencies) are able to provide an observable benefit.

To assess this threshold hypothesis and the fluency hypothesis proposed by Peters et al. (2006), Hill and Brase (2012) systematically tested a variety of problem types with varying levels of difficulty and in different number formats, while also assessing numerical literacy with the standard measure used in this research (i.e., the General Numeracy Scale; Lipkus et al., 2001). These findings generally showed an absence of any interaction across several different problem types. Of most importance to the current paper, the Bayesian reasoning problems originally used by Chapman and Liu (2009) also failed to replicate the numeracy × number format interaction, causing some specific concern over the "threshold hypothesis" of Bayesian reasoning, and to a lesser extent the "fluency hypothesis" of judgment and decision making tasks in general. The one constant across these studies was a consistent main effect for numeracy and a consistent main effect for number format, with higher numerates performing better on Bayesian reasoning tasks, and participants given the natural frequencies format also performing better than those given single event probability versions.

Support for the findings of Hill and Brase (2012) were shown by Garcia-Retamero and Hoffrage (2013) who studied the Bayesian reasoning ability of doctors and patients in medical decision tasks. After fully crossing conditions by number format (natural frequencies and single event probabilities) and display (number only or pictorial representation), participants' numeracy scores were also assessed. Garcia-Retamero and Hoffrage (2013) found the traditional frequency effect, just as in Hill and Brase (2012), and also an improvement in Bayesian reasoning performance by including a pictorial representation. Numeracy did not interact with the frequency effect, again consistent with the Hill and Brase (2012) findings and with the ecological rationality explanation of the frequency effect. Johnson and Tubau (2013) also partially replicated the lack of a numeracy ×

number format interaction, and found consistent improvement in Bayesian reasoning as a result of using natural frequencies, with the only exception being in very difficult problems, operationally defined by longer word length of the problem text. Johnson and Tubau (2013) proposed that both Chapman and Liu (2009) and Hill and Brase (2012) may be partially correct. When given long ("difficult") problems, the numeracy × number format interaction was present, with low numerates showing a floor effect, and high numerates showing the benefit of natural frequencies, a finding consistent with the "threshold hypothesis" of Chapman and Liu (2009). However, with less difficult problems the numeracy × number format interaction disappeared, a finding in line with Hill and Brase (2012).

The above set of results led Johnson and Tubau (2013) to suggest a potential problem with evolutionary accounts proposed by various researchers (e.g., Cosmides and Tooby, 1996; Brase et al., 1998), in that there was not a frequency facilitation effect for the *very* difficult problems. The present authors, however, do not see this as a problem for an evolutionary account. We reach this conclusion because differences in *problem context* (e.g., problem difficulty, word count) that are assessed in terms of the written problem properties are only tenuously connected to evolved cognitive abilities. Cognitive mechanisms evolved to solve specific problems in specific environments. The perspective of ecological rationality, which is generally consistent with evolutionary psychology, is also built upon a similar premise (i.e., the fit between the structure of the environment and the design of the mind; Gigerenzer et al., 1999; Gigerenzer and Gaissmaier, 2011). By analogy, this situation can be compared to someone proposing that humans have an evolved ability to develop complex language. This proposal is not endangered by the observation that people (even highly literate people) find a college physics textbook difficult to read. Reading is a cultural invention which taps into our evolved language ability, and thus our ability to handle a particularly difficult written text is only tenuously connected to the evolved cognitive ability for human language.

More recent work on individual difference moderators of the frequency effect in Bayesian reasoning has only made the aforementioned research more perplexing. For instance, McNair and Feeney (2015) demonstrated a "threshold" type effect despite slightly different problem format manipulations. Specifically, McNair and Feeney (2015) assessed the differences between the standard format (single event probabilities) and a causal format (still single event probabilities, but with additional text describing a possible cause for false positive test results); previous research by Krynski and Tenenbaum (2007) demonstrated evidence that causal structures in problems could lead to improved Bayesian reasoning performance. In separate studies, McNair and Feeney (2015) found evidence for numerical literacy serving as a moderator of problem structure's benefits on Bayesian accuracy, with the effect of problem structure only present in highly numerate individuals. Similar to the discussion of the threshold hypothesis of Chapman and Liu (2009), this observation of an apparent moderating relationship between privileged representational formats, and individual difference measures (e.g., numeracy, cognitive reflection) might be seen as damaging to evolutionary and ecological accounts. However, the same explanation as offered for the Chapman and Liu (2009) results can hold for the McNair and Feeney (2015) results: that performance near floor effect levels can resemble an interaction. In fact, performance in the McNair and Feeney (2015) studies *was* somewhat low (range: 3 to 32% in lowest to highest performing conditions).

Other recent research (Lesage et al., 2013; Sirota et al., 2014a) has addressed a commonly held assumption critics make about the "ecological rationality account": if naturally sampled frequencies are a privileged representational format for an evolved statistical reasoning module, then the module must be "closed," and automatic. Thus, any general cognitive traits (e.g., cognitive reflection), or any method of decreasing general cognitive capacity (e.g., cognitive load), should not significantly interfere with Bayesian performance, or the frequency effect. In general terms, this idea is the assumption of modular encapsulation (Fodor, 1983), which is still promoted by Fodor but actually not accepted by any prominent evolutionary psychology views (e.g., compare Fodor, 2000 and Barrett, 2005).

Although both groups of authors readily acknowledge the research conducted, and the reviews published, concerning the massive modularity hypothesis, there does seem to be some misunderstanding. For example, Barrett and Kurzban (2006, see specifically pp. 636–637), which is cited by some of the work mentioned above, discuss at length the misunderstandings about automaticity of evolved modules, and the method of using cognitive load induced deficits as evidence against evolved modules. Without getting too detailed, their arguments can be summarized by the following analogy: personal computers have a variety of specialized programs (modules). Few would argue that a word processor works as efficiently at storing and computing numerical data, as compared to a spreadsheet program. Thus, these programs are separate, and specialized. However, if I download 1,000 music files to my computer, the overall performance of those separate programs will suffer, at least with respect to processing time. Also, if I drain the battery power in my laptop, the programs will fail to operate at all. This observation does not lead directly to the conclusion that the programs are not specialized. It simply points to the conclusion that the programs require some overlapping general resources. The same conclusion should be made with respect to cognitive modules. The examples in this analogy are extreme instances of general situations which can impair the functioning of functionally specific modules, but the point holds. The question becomes not one of modular abilities being impervious to general resource constraints, but rather one of understanding *how* particular situational contexts influence the functioning of specific cognitive abilities.

In a different study of individual differences, Kellen et al. (2013) found the standard benefits of pictorial representations (Venn diagrams, in this case) in answering complex statistical tasks such as Bayesian reasoning. Furthermore, this general pattern interacted with measured spatial ability, which was independently assessed. In low-complexity problems, low spatial ability participants actually were hurt by pictorial representations, whereas high spatial ability participants demonstrated no difference between pictorial and text displays. However, in

high-complexity problems, high spatial ability participants were aided in their understanding by the presence of pictorial representations, whereas low spatial ability participants saw no benefit. This last result is somewhat consistent with a threshold hypothesis, but there are many issues within these studies in need of deeper assessment. Further research is needed to clarify how different spatial ability levels are related to the use of different types of visual displays and if there is any relationship between spatial ability, numeracy, and the effects of naturally sampled frequencies.

Finally, there are differences in performance that are related to the incentive structures under which people are asked to do Bayesian reasoning tasks. Research participants who do Bayesian reasoning tasks as part of a college course (either through a research "subject pool" or as in-class volunteers) tend to perform less well than participants who are paid money for their participation (Brase et al., 2006). This same research also documented that participants from more selective universities generally performed better than those from less selective universities, most likely due to a combination of different overall ability levels and different intrinsic motivation levels to do academic-type tasks. Brase (2009b) extended this research to show that people whose payments were tied to performance (i.e., correct responses received more money) did even better than people who were given a flat payment for their participation. This is an important factor in, for example, understanding the very high level of Bayesian reasoning performance found by Cosmides and Tooby (1996; paid participants from Stanford University) versus the lower performance on the same task in Sloman et al. (2003; in-class participants from Brown University). In all cases, however, it should be noted that the relative levels of performance when varying the use of natural sampling, frequencies, and pictorial aids were consistent across studies. Absolute performance levels vary, but these methods for improving Bayesian reasoning remain effective.

Conclusion

Overall, the literature on Bayesian reasoning is clear and straightforward in terms of *what* works for improving performance: natural sampling, frequencies, icon-based pictures, and more general development of the prerequisite skills for these tasks (i.e., numerical literacy, visual ability, and motivation to reach the correct answer). The more contentious topic is that of *why* these factors work to improve Bayesian reasoning. The balance of evidence favors the ecological and evolutionary rationality explanations for why these factors are key to improving Bayesian reasoning. This verdict is supported by multiple considerations which flow from the preceding review. First, the ecological rationality account is consistent with a broad array of scientific knowledge from animal foraging, evolutionary biology, developmental psychology, and other areas of psychological inquiry. Second, the ecological rationality approach is the view which has consistently tended to discover and refine the existence of these factors based on *a priori* theoretical considerations, whereas alternative accounts have tended to emerge as *post hoc* explanations. (To be specific, the facilitation effect of natural frequencies documented by Gigerenzer and Hoffrage (1995), the facilitative effect of pictorial representation documented by Cosmides and Tooby (1996), the effect of using whole objects versus aspects of objects documented by Brase et al. (1998), and the differential effects of specific types of pictorial aids in Bayesian reasoning documented by Brase (2009a, 2014) all were established based on ecological rationality considerations which were then followed by alternative accounts.) Third, the actual nature of the evidence itself supports the ecological rationality approach more than other accounts. For instance, in head-to-head evaluations of rival hypotheses, using uncontestable methodologies, the results have supported the ecological rationality explanations (e.g., Brase, 2009a). Furthermore, a quite recent meta-analysis (McDowell and Jacobs, 2014) has conclusively established the validity of the effect of naturally sampled frequencies in facilitating Bayesian reasoning, as described from an ecological rationality perspective.

Distressingly, some proponents of a heuristics and biases view of Bayesian reasoning have not engaged with the bulk of the above literature which critically evaluates this view relative to the ecological rationality view. As just one illustration, Ayal and Beyth-Marom (2014) cite the seminal work by Gigerenzer and Hoffrage (1995), yet ignore nearly all of the other research done from an ecological rationality approach in the subsequent nearly 20 years. Robert Frost (1919/1999) noted that people often say "good fences make good neighbors," but that this is not necessarily a true statement:

> *Before I built a wall I'd ask to know*
> *What I was walling in or walling out,*
> *And to whom I was like to give offence.*
> *Something there is that doesn't love a wall,*

In science, perhaps even more than in other domains of life, fences are *not* good. Willingness to engage openly, honestly, and consistently with the ideas one does not agree with should be a hallmark of scientific inquiry. Failing to do so is scientifically irresponsible.

In conclusion, the vast majority of studies in human Bayesian reasoning align well with evolutionary and ecological rationality account of how the mind may be designed. These accounts are theoretically parsimonious and established in a rich set of literature from a wide range of interrelated disciplines. Alternative explanations, however, tend to appeal to stripped down parts of this account, often losing clear predictive power in the process, which neglect the ecological and evolutionary circumstances of the human mind they purport to explain. That does not mean that the heuristic and biases account no longer has any validity. The intellectually invigorating component of this debate is that we do not fully understand all that is to learn about how people engage in (or fail to engage in) Bayesian reasoning. There is still much to learn about the possible environmental constraints on Bayesian reasoning (e.g., problem difficulty, number of cues), and how those constraints may be interwoven with individual differences (e.g., numerical literacy, spatial ability), and even different measures of specific individual differences (e.g., subjective vs. objective numeracy). We look forward to disassembling walls

and integrating various perspectives, with the hope of more fully understanding how to improve Bayesian reasoning, and how those methods of improvement illuminate the nature of human cognition.

Acknowledgment

Publication of this article was funded in part by the Kansas State University Open Access Publishing Fund.

References

Ayal, S., and Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgm. Decis. Mak.* 9, 226–242.

Barrett, H. C. (2005). Enzymatic computation and cognitive modularity. *Mind Lang.* 20, 259–287. doi: 10.1111/j.0268-1064.2005.00285.x

Barrett, H. C., and Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychol. Rev.* 113, 628–647. doi: 10.1037/0033-295X.113.3.628

Bramwell, R., West, H., and Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: experimental study. *Br. Med. J.* 333, 284–289. doi: 10.1136/bmj.38884.663102.AE

Brase, G. L. (2002). Ecological and evolutionary validity: comments on Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni's (1999) mental model theory of extensional reasoning. *Psychol. Rev.* 109, 722–728. doi: 10.1037/0033-295X.109.4.722

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychon. Bull. Rev.* 15, 284–289. doi: 10.3758/PBR.15.2.284

Brase, G. L. (2009a). Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol.* 23, 369–381. doi: 10.1002/acp.1460

Brase, G. L. (2009b). How different types of participant payoffs alter task performance. *Judgm. Decis. Mak.* 4, 419–428.

Brase, G. L. (2014). The power of representation and interpretation: doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *J. Cogn. Psychol.* 26, 81–97. doi: 10.1080/20445911.2013.861840

Brase, G. L., Cosmides, L., and Tooby, J. (1998). Individuation, counting, and statistical inference: the roles of frequency and whole object representations in judgment under uncertainty. *J. Exp. Psychol. Gen.* 127, 3–21. doi: 10.1037/0096-3445.127.1.3

Brase, G. L., Fiddick, L., and Harries, C. (2006). Participant recruitment methods and statistical reasoning performance. *Q. J. Exp. Psychol.* 59, 965–976. doi: 10.1080/02724980543000132

Casscells, W., Schoenberger, A., and Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. *N. Engl. J. Med.* 299, 999–1000. doi: 10.1056/NEJM197811022991808

Chapman, G. B., and Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgm. Decis. Mak.* 4, 34–40.

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

Edwards, W. (1982). "Conservatism in human information processing," in *Judgment Under Uncertainty: Heuristics and Biases,* eds D. Kahneman, P. Slovic, and A. Tversky (New York, NY: Cambridge University Press).

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

Fodor, J. (1983). *The Modularity of Mind: An Essay in Faculty Psychology*. Cambridge, MA: MIT Press.

Fodor, J. A. (2000). *The Mind Doesn't Work that Way: The Scope and Limits of Computational Psychology*. Cambridge, MA: The MIT Press.

Frost, R. (1919/1999). "Mending Wall," in *Modern American Poetry: An Introduction,* ed. L. Untermeyer (New York, NY: Harcourt, Brace and Howe), Available at: www.bartleby.com/104/

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

Gigerenzer, G. (2001). Content-blind norms, no norms, or good norms? A reply to Vranas. *Cognition* 81, 93–103. doi: 10.1016/S0010-0277(00)00135-9

Gigerenzer, G., and Gaissmaier, W. (2011). Heuristic decision making. *Annu. Rev. Psychol.* 62, 451–482. doi: 10.1146/annurev-psych-120709-145346

Gigerenzer, G., Gaissmaier, W., Kurz-milcke, E., Schwartz, L. M., and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychol. Sci. Public Interest* 8, 53–96. doi: 10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., and Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev.* 106, 425–430. doi: 10.1037//0033-295X.106.2.425

Gigerenzer, G., and Hoffrage, U. (2007). The role of representation in Bayesian reasoning: correcting common misconceptions. *Behav. Brain Sci.* 30, 264. doi: 10.1017/S0140525X07001756

Gigerenzer, G., Todd, P. M., and the ABC Research Group. (1999). *Simple Heuristics that Make us Smart*. New York, NY: Oxford University Press.

Girotto, V., and Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition* 78, 247–276. doi: 10.1016/S0010-0277(00)00133-5

Girotto, V., and Gonzalez, M. (2002). Chances and frequencies in probabilistic reasoning: rejoinder to Hoffrage, Gigerenzer, Krauss and Martignon. *Cognition* 84, 353–359. doi: 10.1016/S0010-0277(02)00051-3

Hill, W. T., and Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *Q J. Exp. Psychol.* 65, 2343–2368. doi: 10.1080/17470218.2012.687004

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition* 84, 343–352. doi: 10.1016/S0010-0277(02)00050-1

Johnson, E. D., and Tubau, E. (2013). Words, numbers, & numeracy: diminishing individual differences in Bayesian reasoning. *Learn Individ. Differ.* 28, 34–40. doi: 10.1016/j.lindif.2013.09.004

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., and Caverni, J.-P. P. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychol. Rev.* 106, 62–88. doi: 10.1037/0033-295X.106.1.62

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognit. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Kellen, V., Chan, S., and Fang, X. (2013). "Improving user performance in conditional probability problems with computer-generated diagrams," in *Human-Computer Interaction: Users and Contexts of Use*, ed. M. Kurosu (Berlin: Springer Berlin Heidelberg), 183–192.

Kleiter, G. (1994). "Natural sampling: rationality without base rates," in *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, eds G. H. Fischer and D. Laming (New York, NY: Springer), 375–388. doi: 10.1007/978-1-4612-4308-3_27

Krynski, T. R., and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *J. Exp. Psychol. Gen.* 136, 430–450. doi: 10.1037/0096-3445.136.3.430

Lesage, E., Navarrete, G., and De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: the role of general cognitive resources. *Think. Reason.* 19, 27–53. doi: 10.1080/13546783.2012.713177

Lewis, C., and Keren, G. (1999). On the difficulties underlying Bayesian reasoning: a comment on Gigerenzer and Hoffrage. *Psychol. Rev.* 106, 411–416. doi: 10.1037/0033-295X.106.2.411

Lipkus, I. M., Samsa, G., and Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Med. Decis. Making* 21, 37–44. doi: 10.1177/0272989X0102100105

Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. *Q. J. Exp. Psychol. Human Exp. Psychol.* 48A, 188–207. doi: 10.1080/14640749508401384

Macchi, L., and Mosconi, G. (1998). Computational features vs frequentist phrasing in the base-rate fallacy. *Swiss J. Psychol.* 57, 79–85.

McDowell, M. E., and Jacobs, P. L. (2014). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Poster Presented at the Society for Judgment and Decision Making Conference*, Long Beach, CA.

McNair, S., and Feeney, A. (2015). Whose statistical reasoning is facilitated by causal structure intervention? *Psychon. Bull. Rev*. 22, 258–264. doi: 10.3758/s13423-014-0645-y

Mellers, B. A., and McGraw, A. P. (1999). How to improve Bayesian reasoning: comment on Gigerenzer and Hoffrage 1995. *Psychol. Rev*. 106, 417–424. doi: 10.1037/0033-295X.106.2.417

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., and Dickert, S. (2006). Numeracy and decision making. *Psychol. Sci*. 17, 407–413. doi: 10.1111/j.1467-9280.2006.01720.x

Sirota, M., Juanchich, M., and Hagmayer, Y. (2014a). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychon. Bull. Rev*. 21, 198–204. doi: 10.3758/s13423-013-0464-6

Sirota, M, Kostovičová, L., and Juanchich, M. (2014b). The effect of iconicity of visual displays on statistical reasoning: evidence in favor of the null hypothesis. *Psychon. Bull. Rev*. 21, 961–968. doi: 10.3758/s13423-013-0555-4

Sloman, S. a., Over, D., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Process* 91, 296–309. doi: 10.1016/S0749-5978(03)00021-9

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

Tversky, A., and Kahneman, D. (1982). "Evidential impact of base rates," in *Judgment Under Uncertainty: Heuristics and Biases*, eds D. Kahneman, P. Slovic, and A. Tversky (Cambridge: Cambridge University Press), 153–160. doi: 10.1017/CBO9780511809477.011

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev*. 90, 293–315. doi: 10.1037/0033-295X.90.4.293

Vranas, P. B. M. (2000). Gigerenzer's normative critique of Kahneman and Tversky. *Cognition* 76, 179–193. doi: 10.1016/S0010-0277(99)00084-0

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: frequency or nested sets? *Exp. Psychol*. 50, 97–106. doi: 10.1026//1618-3169.50.2.97

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

REVIEW
published: 27 July 2015
doi: 10.3389/fpsyg.2015.00938

Comprehension and computation in Bayesian problem solving

Eric D. Johnson[1,2]* *and Elisabet Tubau*[1,2]

[1] Department of Basic Psychology, University of Barcelona, Barcelona, Spain, [2] Research Institute for Brain, Cognition, and Behavior (IR3C), Barcelona, Spain

Humans have long been characterized as poor probabilistic reasoners when presented with explicit numerical information. Bayesian word problems provide a well-known example of this, where even highly educated and cognitively skilled individuals fail to adhere to mathematical norms. It is widely agreed that natural frequencies can facilitate Bayesian inferences relative to normalized formats (e.g., probabilities, percentages), both by clarifying logical set-subset relations and by simplifying numerical calculations. Nevertheless, between-study performance on "transparent" Bayesian problems varies widely, and generally remains rather unimpressive. We suggest there has been an over-focus on this representational facilitator (i.e., transparent problem structures) at the expense of the specific logical and numerical processing requirements and the corresponding individual abilities and skills necessary for providing Bayesian-like output given specific verbal and numerical input. We further suggest that understanding this task-individual pair could benefit from considerations from the literature on mathematical cognition, which emphasizes text comprehension and problem solving, along with contributions of online executive working memory, metacognitive regulation, and relevant stored knowledge and skills. We conclude by offering avenues for future research aimed at identifying the stages in problem solving at which correct vs. incorrect reasoners depart, and how individual differences might influence this time point.

Keywords: Bayesian reasoning, mathematical problem solving, text comprehension, set-subset reasoning, numeracy, individual differences

OPEN ACCESS

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
Jean Baratgin, Université Paris 8, France
Ulrich Hoffrage, Université de Lausanne, Switzerland

***Correspondence:**
Eric D. Johnson, Departament de Psicologia Bàsica, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron 171, 08035 Barcelona, Spain
eric.johnson@ub.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *30 March 2015*
Accepted: *22 June 2015*
Published: *27 July 2015*

Citation:
Johnson ED and Tubau E (2015) Comprehension and computation in Bayesian problem solving. Front. Psychol. 6:938. doi: 10.3389/fpsyg.2015.00938

Introduction

Over the past decades, there has been a growing appreciation for the probabilistic operations of human cognition. The union of highly sophisticated modeling techniques and theoretical perspectives, sometimes referred to as the "Bayesian Revolution," is posed to bridge many traditional problems of human inductive learning and reasoning (Wolpert and Ghahramani, 2005; Chater and Oaksford, 2008; Tenenbaum et al., 2011). Despite this promising avenue, probabilistic models have acknowledged limits. One of the most prominent of these is the persistent difficulties that even highly educated adults have reasoning in a Bayesian-like manner with explicit statistical information (Kahneman and Tversky, 1972; Gigerenzer and Hoffrage, 1995; Barbey and Sloman, 2007), including individuals with advanced education (Casscells et al., 1978; Cosmides and Tooby, 1996), higher cognitive capacity (Lesage et al., 2013; Sirota et al., 2014a), and higher numeracy skills (e.g., Chapman and Liu, 2009; Hill and Brase, 2012; Johnson and Tubau, 2013; Ayal and Beyth-Marom, 2014; McNair and Feeney, 2015). Rather than contradicting Bayesian models of reasoning, however, less than optimal inferences over explicit verbal and numerical

information result in large part from the relatively recent cultural developments of these symbolic systems, far too little time for evolution to have automated this explicit reasoning capacity.

In the present review, we focus on Bayesian word problems, or the *textbook-problem paradigm* (Bar-Hillel, 1983), where a binary hypothesis and observation (e.g., the presence of a disease, the results of a test) are verbally categorized and numerically quantified within a hypothetical scenario. We use the term "Bayesian word problems" to refer to tasks in which these explicitly summarized statistics are provided as potential input for a Bayesian inference in order to derive a posterior probability (these correspond to "statistical inference" tasks in Mandel, 2014a). Specifically, the base rate information (e.g., the probability of having a disease) has to be integrated with the likelihood of a certain observation (e.g., the validity of a diagnostic test, reflected in a hit rate, and false-positive rate) to arrive at precisely quantified Bayesian response (e.g., the probability of having the disease conditioned on a positive test). Hence, these problems reflect situations of *focusing* (rather than *updating per se*), where an initial state of knowledge is refined, or re-focused, in an otherwise stable universe of possibilities (Dubois and Prade, 1992, 1997; Baratgin and Politzer, 2006, 2010). Given this static coherence criterion of these word problems, the normative view of additive probability theory holds (Kolmogorov, 1950), and so Bayes' rule is the most appropriate normative standard for assessing performance (see Baratgin, 2002; Baratgin and Politzer, 2006, 2010)[1].

Some have argued that Bayesian word problems may in fact have little to do with "Bayesian reasoning" in the sense that they do not necessarily require updating a previous belief (see Koehler, 1996; Evans et al., 2000; Girotto and Gonzalez, 2001, 2007; Mandel, 2014a; Girotto and Pighin, 2015). This sentiment reflects a gradual shift from using these tasks to understand how well (or poorly) humans update the probability of a hypothesis in light of new evidence, or how experienced physicians diagnose disease given a specific indicator (Casscells et al., 1978; Eddy, 1982), to the task features and individual differences associated with reasoning outcomes, which are often found to depart from the Bayesian ideal (Barbey and Sloman, 2007; Navarrete and Santamaría, 2011; Mandel, 2014a). We take this descriptive-normative gap to be our general question: *Why do people tend to deviate, often systematically, from the normative standard prescribed by Bayes' rule*?

Fortunately not all is lost, and a variety of factors are increasingly understood which can be manipulated to facilitate Bayesian responses from floor to near ceiling performance. In what follows, we first aim to clarify some frequently confused terms, isolate key factors influencing performance, and point out some limitations of typically contrasted theoretical views. We then highlight some mutually informative parallels between research and theory on Bayesian inference tasks, and the literature on mathematical problem solving and education. Finally, we discuss how these separate, but complimentary, views on reasoning and mathematical cognition can provide some general processing considerations and new methodologies relevant for understanding why human performance falls short of Bayesian ideals, and how this gap might be reduced.

[1]In this review we do not address the distinction between logical and subjective Bayesianism, nor do we refer to situations involving a dynamic cohesion criterion or the conditioning principle (see Baratgin and Politzer, 2006) in which other normative standards may apply (for discussion on the normative issue see Gigerenzer, 1991; Koehler, 1996; Vranas, 2000; Baratgin and Politzer, 2006, 2010; Douven and Schupbach, 2015).

Natural Frequencies: from Base-rate Neglect to Nested-sets Respect

In the present section we explore the Bayesian reasoning task, using a variant of the classic medical diagnosis problem (Casscells et al., 1978; Eddy, 1982) as a general point of reference. We center on the natural frequency effect—a facilitator of both representation *and* computation—and the debate which has surrounded it for nearly two decades. We highlight the general consensus on the benefits of making nested-set structures transparent, before turning to other processing requirements needed for transforming presented words and numbers into a posterior Bayesian response in the following section.

Poor Reasoning and Base-rate Neglect

> "In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all."
>
> Kahneman and Tversky (1972 p. 450)

Although Bayesian norms have been around since the 18th century (Bayes, 1764), it was not until 200 years later that psychological research adopted these standard as the benchmark against which to measure human reasoning ability. As exemplified in the quote above, early results were not too promising. In the heyday of the heuristics-and-biases paradigm, one of medicine's most coveted journals, the *New England Journal of Medicine*, published a study where a group of medically trained physicians were given the following problem (Casscells et al., 1978):

> If a test to detect a disease whose prevalence is 1/1,000 [**BR**] has a false positive rate of 5% [**FPR**], what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs?___%[2].

This medical diagnosis problem asks for the probability (chance) that a person actually has the disease (the hypothesis) given a positive test result (the data), a task of which physicians should be reasonable adept. The results, however, were not very encouraging, with only 18% of the physicians answering with the Bayesian response of 2%. Forty-five percent of them, on the other hand, answered "95%," which appeared to completely ignore the base rate presented in the problem—the fact that only 1 in 1000 people actually have the disease. Similar results

[2]Information in [brackets] was not present in original text, but is included in examples in this review to ease cross-problem comparisons. [BR] = base rate, [FPR] = false-positive rate. Implicit in this example is the hit rate [HR] = 1.

were reported a few years later by Eddy (1982). Evidence was accordingly interpreted to show that humans tend to neglect crucial information (such as base rates), while instead focusing on the similarity of target data to prototypical members of a parent category (for reviews see Kahneman et al., 1982; Koehler, 1996). This was part of a larger explanatory framework which emphasized limited cognitive processing capacity, where mental shortcuts, or heuristics, are employed to alleviate the burden of cognitively demanding tasks, including those that may be more optimally answered with formal calculations (e.g., Kahneman, 2003). However, "base-rate neglect" as a general explanation has been critiqued on theoretical and methodological grounds (Koehler, 1996), which is further supported by the observation that typical errors in Bayesian word problems tend to be a function of the question format, with base-rate-only responses often reported (e.g., Gigerenzer and Hoffrage, 1995; Mellers and McGraw, 1999; Evans et al., 2000; Girotto and Gonzalez, 2001).

The Natural Frequency Effect: Evolution and Computation

At a time when pessimism dominated the landscape of the cognitive psychology of reasoning, Gigerenzer and Hoffrage (1995) and Cosmides and Tooby (1996) offered hope for the human as statistician, along with a strong theoretical agenda (see also Brase et al., 1998). Consider this *frequency* alternative to the Casscells et al. (1978) medical diagnosis problem presented above:

> 1 out of every 1000 [**BR**] Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive [**HR=1**]. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive [**FPR**] for the disease.
> Imagine that we have assembled a random sample of 1000 Americans. They were selected by a lottery. Those who conducted the lottery had no information about the health status of any of these people. Given the information above, on average, how many people who test positive for the disease will actually have the disease? __ out of __.

Performance on this problem was found to elicit a correct response rate of 72% by Cosmides and Tooby (1996, study 2), remarkably higher than the 18% reported by Casscells et al. with the formally analogous information shown above. In a similar vein, Gigerenzer and Hoffrage (1995, 1999) reported success rates near 50% across a variety of problems presenting *natural frequencies*, compared to 16% with their probability versions. Examples of similar problems presenting natural frequencies and normalized data are shown in **Figure 1**.

The initial explanations offered for these effects can be divided in two strands: *Evolution* and *computation*. According to Cosmides and Tooby, evolution endowed the human mind with a specialized, automatically-operating frequency module for making inferences over countable sets of objects and events, but which is ineffective for computing single-event probabilities. By tapping into this module naïve reasoners can solve frequency problems, while they fail on probability problems because this module cannot be utilized. Relatedly, Gigerenzer and Hoffrage suggested that reasoning performance depended on the mesh between the presented problem data (the structure of the task environment) and phylogenetically endowed cognitive algorithms for *naturally sampled* information (Kleiter, 1994; **Figure 2B**), which leads to a similar suggestion that explicit numerical reasoning would utilize the same cognitive processes used for reasoning based on information experienced over time, provided the external input matched the internal algorithm. Unlike Cosmides and Tooby (1996) and Gigerenzer and Hoffrage (1995) did not specifically argue that the mind is unable to deal with probabilities of single events, and in fact their computational account predicted quite the opposite (see their study 2 and "prediction 4").

The more pertinent claim of Gigerenzer and Hoffrage (1995), however, was their *computational* analysis, which focused on the difference between the information provided in the problem and its proximity to the Bayesian solution. With normalized information (e.g., percentages; see **Table 1**), the following

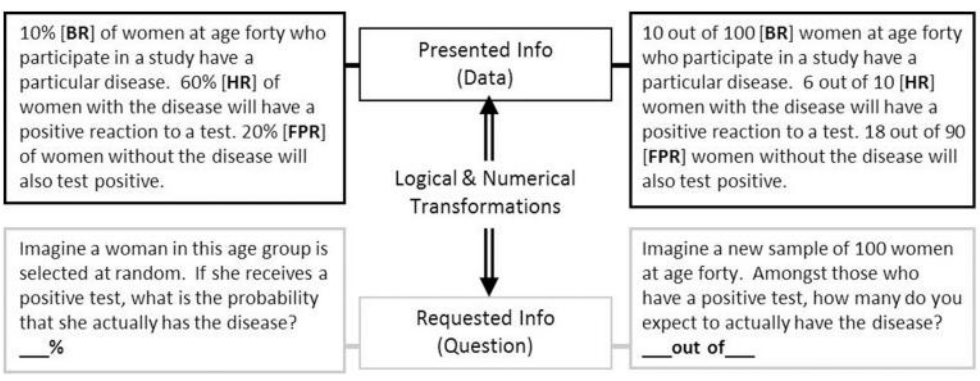

FIGURE 1 | Examples of the medical diagnosis problem, presented with *normalized* numerical information (left) and with *natural frequencies* (right). If not otherwise indicated, other tables, figures, and examples in the text refer to the numerical information in this figure.

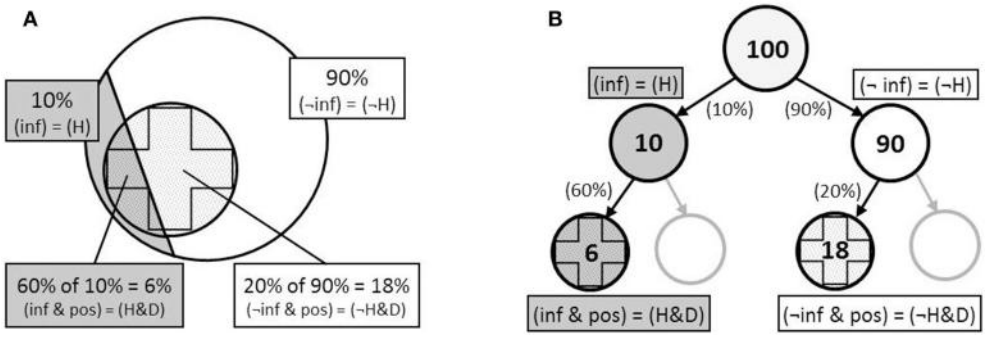

FIGURE 2 | Representations of the Bayesian tasks presented in Figure 1 as (A) integrated nested sets, and (B) frequency tree, where (*H*) the hypothesis = infected (inf); and (*D*), the data = positive test (pos).

computation is necessary to arrive at a Bayesian response, where H is the hypothesis (having the disease) and D is the data (testing positive):

$$p(H|D) = \frac{p\,(H\&D)}{p(D)} = \frac{p(H)p(D|H)}{p(H)p(D|H) + p(\neg H)p(D|\neg H)}$$
$$= \frac{(0.1)(0.6)}{(0.1)(0.6) + (0.9)(0.2)} = \text{“25\%”}$$

With natural frequencies, on the other hand, all numerical information is absolutely quantified to a single reference class (namely, the superordinate set of the problem; “100 people” in **Figure 1**; see also **Figure 2B**), where categories are naturally classified into the joint occurrences found in a bivariate 2 × 2 table [e.g., ($H\&D$), ($\neg H\&D$)]. In this case, the conditional distribution does not depend on the between-group (infected, not infected) base rates, but only on the within-group frequencies (hit rate, false-positive rate). Accordingly, base rates can be ignored, numbers are on the same scale and can be directly compared and (additively) integrated, and the required computations are reduced to a simpler form of Bayes rule:

$$p(H|D) = \frac{p\,(H\&D)}{p(D)} = \frac{p(H\&D)}{p(H\&D) + p(\neg H\&D)}$$
$$= \frac{6}{6+18} = \text{“6 } out\ of \text{ 24”}$$

Thinking in Sets: Comprehension and Manipulation of Nested-set Structures

While the computational simplification afforded by natural frequencies was clear, critiques of the evolutionary view came quickly. Over the ensuing decade, a number of studies appeared which argued that the “frequency advantage” was better described as a “structural advantage” (Macchi, 1995; Macchi and Mosconi, 1998; Johnson-Laird et al., 1999; Lewis and Keren, 1999; Mellers and McGraw, 1999; Evans et al., 2000; Girotto and Gonzalez, 2001, 2002; Sloman et al., 2003; Yamagishi, 2003; Fox and Levav, 2004). More specifically, these studies suggested that the benefit of natural frequencies was not in the numerical format *per se* (frequencies vs. percentages), nor the number of events being reasoned about (sets of individuals vs. single-event probabilities), but rather in the clarification of the abstract nested-set relationships inherent in the problem data, which helps reasoners to form appropriate models of relevant information. The nested-set structure of these problems is illustrated in **Figure 2**, where it can be seen that the relations between categories—people *infected* (H) and *not infected* ($\neg H$) testing *positive* (D) for a disease—can be represented spatially as a hierarchical series of nested sets.

Clearly, the quantitative relationships amongst subsets are more transparently afforded with natural frequencies compared to normalized percentages. This general view emphasizing *representational* facilitation has come to be known as the *nested-sets hypothesis*, originally proposed by Tversky and Kahneman (1983), and which has since been variously expressed by a number of authors. For example, Mellers and McGraw (1999) concluded that natural frequencies are “advantageous because they help people visualize nested sets, or subsets relative to larger sets” (p. 419). Girotto and Gonzalez (2001) attributed successful reasoning to problem presentations which “activate intuitive principles based on subset relations” (p. 247). For Evans (2008), “what facilitates Bayesian reasoning is a problem structure which cues explicit mental models of nested-set relations” (p. 267). And as stated by Barbey and Sloman (2007, p. 252): “the mind embodies a domain general capacity to perform elementary set operations and that these operations can be induced by cues to the set structure of the problem.” Although these suggestions are not without limitations (discussed below), proponents of the nested-sets hypothesis helped identify a key strategy that reasoners (naïve to Bayes rule) *can* use to arrive at a Bayesian response: *Thinking in sets*. That is, in the absence of formal knowledge of how to optimally combine conditional probabilities, reasoners can still solve these tasks by considering the problem as overlapping sets of data, namely, as a focal subset of infected people out of the reference set of people who test positive: ($H\&D$)/(D).

Contemporary discussions explaining Bayesian facilitations continue to be framed in terms of a nested-sets (or domain general) contra an ecological rationality (or frequency/format-specific) debate (e.g., Navarrete and Santamaría, 2011; Hill and Brase, 2012; Lesage et al., 2013; Brase, 2014; Sirota et al., 2014a,b,

2015a,b; Brase and Hill, 2015). We assume that theorists on both sides of the divide are more interested in finding out how to improve Bayesian reasoning and why these facilitations work, rather than simply promoting a preferred position. We also believe that, in general, these perspectives may in some regards be more complimentary than adversarial. Accordingly, we think that it is important to acknowledge what these views have in common, where the relevant differences between these views lie, and whether either view can fully account for empirical data.

To begin, it should by now be well understood that *natural frequencies* do not simply refer to the use of frequency formats, but essentially refer to problem structure as well (Gigerenzer and Hoffrage, 1999, 2007; Hoffrage et al., 2002; for concurrence see Barbey and Sloman, 2007, response R3). Both the natural frequency and nested-sets views agree that frequencies that do not conform to a natural sampling, or partitive, structure are not much better than percentage formats (Evans et al., 2000; Girotto and Gonzalez, 2001; Sloman et al., 2003). Where these two views primarily diverge is in how comfortable they are making precise predictions based on evolutionary claims. For the moment we suggest putting the evolutionary claims aside, and instead focusing on two points of commonality. First, natural frequencies (or problems presenting a "partitive" or "nested-set" structure, or conforming to the "subset principle"; see Brase and Hill, 2015) are widely agreed to be the most general and robust facilitator of Bayesian-like performance. Second, natural frequencies facilitate both representation *and* computation.

We suggest that in order to advance the discussion, we need to move away from the standard "natural frequency vs. nested-sets" debate and instead consider the processing requirements, and corresponding difficulties, given a particular problem presentation (see also McNair, 2015; Vallée-Tourangeau et al., 2015b). In the following section we note key performance variables and often confused issues, and review available evidence looking separately at problems presenting and requesting normalized vs. natural frequency information.

The Bayesian Problem: from Words and Numbers to Meaningful Structures

Table 1 presents some commonly used terms that are often used in different ways and which frequently lead to confusion. Below we briefly highlight the most frequently confused factors (see also Barton et al., 2007).

First, numerical format and the number of events are fully orthogonal dimensions. Normalized formats (e.g., percentages, decimals) can express single-event probabilities (e.g., "10% chance of infection") or proportions of a set (e.g., "10% of people are infected"), and whole numbers can be used to express frequencies (e.g., 10 of 100 people) or single events (10 of 100 chances). This applies both to the information presented in the text and requested in the question.

Second, the "sampling structure" (also referred to as "information structure" or "menu") refers to the specific categorical-numerical information used to express the hit rate and false-positive rate, and is also orthogonal to the above two distinctions (numerical format, number of events). Typically, this refers to the presentation of the conjunctive/joint events [(H&D) and ($\neg H$&D)] vs. the conditional/normalized data [($D|H$) and ($D|\neg H$), along with the base rates (H) and ($\neg H$)]. Any of these categories can be quantified with either frequencies or normalized formats.

Finally, throughout this review we use the term "natural frequencies" to refer to problems which (1) present *whole numbers* (2) in a *natural sampling* (or *partitive*) structure (specifically, one which directly presents H&D and $\neg H$&D), and (3) request *responses as an integer pair*. We acknowledge that on some accounts natural frequencies may refer only to the initial problem data independent of the question format. However, as we review, the primary benefits of natural frequencies hold only when the question also requests a pair of integers (Ayal and Beyth-Marom, 2014), and therefore for ease of exposition we use *natural frequencies* only when all three conditions are present (unless otherwise stated). In contrast, we refer to "normalized" problems as those which do not meet these three criteria (see **Table 1**).

Why do natural frequencies facilitate Bayesian-like responses? In order to answer this question, we have to understand what was so hard in the first place. That is, a facilitation must always be made relative to some initial point, and it is therefore important first to understand why normalized versions are so difficult. We will then be in a better position to understand the facilitating effects of natural frequencies, and more generally why even clearly presented problems can still be so difficult for many reasoners. In the remainder of this section we therefore review factors that have been shown to facilitate, or impair, Bayesian-like reasoning with problems presenting normalized information or natural frequencies separately.

Reasoning with Normalized Formats

Reasoning with normalized formats is notoriously difficult. However, observing that more "transparent" problems facilitate performance does not necessarily imply that normalized versions are hard simply because the presented data is more difficult to represent. As reviewed below, the difficulty of these problems cannot be reduced to a single (representational or computational) factor. Although some improvements have been observed with visual diagrams and verbal manipulations to the text and question, as well as for individuals with higher cognitive and numerical ability, all of these are limited in their effectiveness.

Visual Representations

Some evidence suggests that visual aids may boost performance with normalized data, which presumably help reasoners to appreciate nested-set relations (for recent reviews see Garcia-Retamero and Cokely, 2013; Garcia-Retamero and Hoffrage, 2013). For example, Sedlmeier and Gigerenzer (2001) showed that training individuals to use frequency trees could have substantial and lasting effects on complex Bayesian reasoning scenarios. Mandel (2015) more recently showed that similar instructions on information structuring improve the accuracy and coherence of probability judgments of intelligence analysts. Recent work by Garcia-Retamero and Hoffrage (2013) also showed substantial benefits of visual aids with probability

TABLE 1 | Key dimensions along which a Bayesian word problem may vary.

Dimension	Description and variables
Numerical format	The format of the presented numerical information: *Whole number integer pairs* (e.g., 10 of 100) vs. *Normalized* (e.g., 10%, 0.1). Formats can be mixed within a single problem.
Question format	The format of the requested response, typically: *Integer pair* (e.g., "__ out of __") vs. *Normalized* (e.g., "__%").
Number of events	*Single-event* (e.g., probability, chance) vs. *Set of events* (e.g., individuals, chances). Can apply to both the presented data ("information type") or to the information requested ("task domain"). Often confused with numerical format, but these are orthogonal issues.
Sampling structure	The particular categorical-numerical information used to express the hit rate and false-positive rate, typically: *Natural* (*H&D*, *¬H&D*; also partitive, transparent, conjunctive, joint) vs. *Normalized* (*D*|*H* and *D*|*¬H*; also non-partitive, relative frequencies, conditional).
Natural frequencies	A problem format which presents *whole numbers* in a *natural sampling* structure (e.g., *H&D*, *¬H&D*), and requests responses as an *integer pair*.
Normalized problems	A problem which presents normalized numerical formats (percentages, decimals), a normalized sampling structure (i.e., with conditional or non-conjunctive information), and/or which requests information in a normalized format (a ratio as a single value, not integer pair).
Context	Scenario of the problem. For example, medical (infection, test); cab (accident, color).
Irrelevant info	Descriptive information that is not relevant for solving the task. Numbers that are not needed for computing the normative response.
Mental steps	The number of steps required to compute the response, given the specific numbers presented in the problem. For example, in **Figure 1**, the number "24" (total positive tests, *D*) is needed but not presented, and must be calculated from $(6 + 18) = 1$ numerical step.
Compatibility	Correspondence between the presented and requested data, including numerical and question formats, also sample sizes.

information. Yamagishi (2003) found that both a roulette-wheel diagram and a frequency tree led to large improvements with information presented as simple fractions (e.g., 1/4, 1/3, 1/2) in the gemstone problem. Sloman et al. (2003) also showed that a Euler circle diagram marginally facilitated performance on a probability version of the medical diagnosis problem. However, in a counterintuitive Bayesian task, the Monty Hall dilemma, Tubau (2008) found no facilitation of a diagrammatic representation of the problem. Overall, while visual diagrams may help with normalized data under some conditions, this facilitation is typically very modest, although instruction or training in information re-representation may be an effective way to improve reasoning in some populations.

Verbal Formulation and Irrelevant Information

There is evidence that reasoning with normalized data can be improved by manipulating the verbal structure of the problem, independent of the numbers provided (Macchi, 1995; Sloman et al., 2003; Krynski and Tenenbaum, 2007; Hattori and Nishida, 2009; Johnson and Tubau, 2013; Sirota et al., 2014a). For example, Macchi (1995) showed how questions which were slightly reformulated to focus on individuating (vs. base-rate) information increased (or reduced) the number of base-rate neglect responses. Sloman et al. (2003, exp. 1) found differences between three numerically identical versions of the medical diagnosis problem, but which varied in the particular wording (or "transparency of nested-set relations") used to transmit the problem data (see also Sirota et al., 2014a, exp. 2). They additionally reported that irrelevant numbers impaired performance, but only with normalized versions (exp. 4B). Johnson and Tubau (2013) also found that simplifying the verbal complexity improved Bayesian outcomes with probabilities, but this was restricted to higher numerate reasoners[3].

Krynski and Tenenbaum (2007) also showed that manipulating verbal content, independent of the numbers, can boost performance. They suggested that reasoners supplement the statistical data presented in the problem with prior world knowledge (of causal relations), and therefore Bayesian reasoning could be enhanced by presenting false-positive rates in terms of alternative causes. Simply providing a cause for the false-positive rate (e.g., "the presence of a benign cyst" in the medical diagnosis context) boosted performance from approximately 25 to 45%, some of the highest performance reported with normalized data in the absence of visual cues. It should be noted, however, that McNair and Feeney (2015; see also 2014) were unable to fully replicate this effect, though they did find evidence that higher numerate reasoners significantly benefitted from a clearer causal structure with normalized information. The participants in Krynski and Tenenbaum's study consisted of undergraduate and graduate students at MIT, who are presumably a more mathematically sophisticated group, which may help to account for the consistent main effect of causal structure in their studies (cf. Brase et al., 2006). This suggests that providing "alternative causes" helped draw attention to the often neglected false-positive data (Evans et al., 2000), which could then be taken advantage of by individuals possessing the requisite numerical skills.

Computation

Normalized versions typically require multiple steps using fraction arithmetic. Despite claims in the reasoning literature that the fraction arithmetic (multiplying and dividing percentages)

[3]*Numeracy* is generally defined as the ability to work with basic numerical concepts, including the comprehension and manipulation of simple statistical and probabilistic information (for reviews see Reyna and Brainerd, 2008; Lipkus and Peters, 2009; Reyna et al., 2009; Peters, 2012). One of the most common measures used in reasoning and decision making studies is the 11-item Lipkus et al. (2001) numeracy scale, which assesses the ability to compare the relative magnitude of ratios, to convert between statistical formats, and to perform simple calculations using frequency ratios and percentages. Other measures of numeracy can be found in, for example, Cokely et al. (2012) and Peters et al. (2007).

required in these tasks in relatively easy (e.g., Johnson-Laird et al., 1999; Sloman et al., 2003), there is indeed substantial evidence that many people lack the requisite conceptual and/or procedural knowledge to correctly carry out these computations (Paulos, 1988; Schoenfeld, 1992; Mayer, 1998; Ni and Zhou, 2005; Reyna and Brainerd, 2007, 2008; Siegler et al., 2011, 2013). Indirect evidence for the difficulty performing multiplicative integrations on normalized problem data is also suggested in Juslin et al. (2011), where it was proposed that reasoners default to less demanding linear additive integrations in the absence of requisite knowledge, cognitive resources, or motivation (see also Juslin, 2015).

An informative result was provided in Ayal and Beyth-Marom (2014). Participants were provided probability information in a percentage format, but the problems were manipulated so that $p(H|D)$ could be computed via a single whole-number subtraction ($1 - p(\neg H|D) = 100 - 92\% = 8\%$). Responses were requested either as a percentage ("compatible") or as frequencies ("incompatible"). In the compatible condition, nearly 80% of higher numerate and around 60% of lower numerate reasoners correctly computed $p(H|D)$; in the incompatible condition, around 70% or of higher numerate and around 34% of lower numerate individuals responded correctly. On the one hand, this expectedly demonstrates that higher numerate individuals are more able to translate between numerical formats. More importantly, however, the high proportion of correct responses, even by reasoners with lower numeracy, demonstrates that the participants in these studies are not inherently unable to understand set relations presented as standardized probabilities. It also shows that the typical computational demands (steps and/or type) with normalized formats may in fact impede Bayesian-like responding (cf. Juslin et al., 2011). It should also be noted, however, that this condition does not require reasoners to understand embedded sets of information (i.e., to simultaneously consider and integrate base rates and diagnostic information); rather, they are simply required to represent the complement of a whole. This implies that the representational difficulty on standard Bayesian problems is not specific to the structure of the data itself, but rather to the relation between the presented and requested information (see also Section Common Processing Demands: Quantitative Backward Reasoning).

Number of Events

Existing research suggests that presenting or requesting single-event probabilities vs. a proportion of a sample (or relative frequencies) with percentages may have little impact on Bayesian responding with normalized data, all else held constant. For example, Gigerenzer and Hoffrage (1995, study 2) found no differences when presenting the data as either relative frequencies with percentages (as in **Figure 1**) vs. single-event probabilities when the question requested a probability. Likewise, Evans et al. (2000, study 2) found no differences with questions requesting a single-event probability vs. a proportion of a sample from data presented as relative frequencies with percentages. While this may be taken as evidence that Bayesian reasoning with percentage information is independent of the number of events referred to, this does not necessarily imply that single-event probabilities are as easily understood as relative frequencies expressed as percentages (e.g., Brase, 2008, 2014; Sirota et al., 2015a; see discussion of "*Chances*" below in Section Reasoning with Natural Frequencies). Recent re-analyses of data from Gigerenzer and Hoffrage (1995) show that problems focusing on individuals (compared to samples, or "numbers") indeed lead to fewer Bayesian responses (Hafenbrädl and Hoffrage, 2015).

Individual Differences

The general finding from individual differences research is that higher cognitive ability, disposition toward analytical thinking, and numeracy level can lead to improved reasoning under some conditions, but to a limited extent (**Table 2**). Sirota et al. (2014a) found that general intelligence (Raven et al., 1977), as well as preference for rational thinking (REI; Pacini and Epstein, 1999), uniquely predicted performance with single-event probabilities. Results of McNair and Feeney (2015) also suggested a significant association between Raven's matrices and performance on normalized Bayesian versions, but an absence of association between the latter and REI. Of note, two studies have reported a lack of association between normalized Bayesian problems and the cognitive reflection test (CRT; Frederick, 2005), a measure of the tendency to suppress initial intuitions and engage in more demanding analytical processing (Lesage et al., 2013; Sirota et al., 2014a). Together, these results suggest that providing the posterior Bayesian ratio with normalized information will necessarily depend on high levels of cognitive ability *and* numeracy. Without these basic requisites, reflective thinking or disposition toward analytical thinking are likely to be of little help (De Neys and Bonnefon, 2013). It is also important to note that even the performance of "higher" ability individuals typically remains quite low. Nevertheless, few studies have directly investigated these factors and results have been mixed, therefore more research is needed to clarify when (and in what combination) individual differences measures are likely to be relevant (proposals of the relative dependencies of these factors can be found in Stanovich, 2009; Klaczynski, 2014; see also Thompson, 2009).

Reasoning with Natural Frequencies

Providing information as natural frequencies (or naturally partitioned sets of chances) is widely hailed as the most effective and robust facilitator of Bayesian-like reasoning. Nevertheless, between-study performance varies widely, and success even with natural frequencies generally remains rather unimpressive (see Newell and Hayes, 2007; Girotto and Pighin, 2015; McNair, 2015). Why do so many individuals still fail to solve these problems even when the structures of these tasks are made "transparent?"

Computation

In their standard form, natural frequencies typically require only a single addition of two whole numbers to construct the needed reference set (D), and the selection of the joint occurrence ($H\&D$) directly provided in the text, to answer the Bayesian question "*(H&D) out of (D).*" Clearly, the whole-number arithmetical demands of the task are manageable by the undergraduate

TABLE 2 | Summary of significant individual differences effects reported in Bayesian word problems presenting normalized information or natural frequencies.

	Numeracy/education	IQ-raven	CRT [I]	Thinking disposition
NORMALIZED VERSIONS*				
Chapman and Liu, 2009	No			
Siegrist and Keller, 2011	Yes/No [a]			
Hill and Brase, 2012	No			
Garcia-Retamero and Hoffrage, 2013	Yes			
Johnson and Tubau, 2013	Yes/No [a]			
Lesage et al., 2013			No	
Sirota et al., 2014a		Yes	No	Yes/No [b]
Ayal and Beyth-Marom, 2014	Yes [c]			
McNair and Feeney, 2015	Yes/No [d]	Yes		No [e]
NATURAL FREQUENCIES				
Brase et al., 2006	Yes			
Chapman and Liu, 2009	Yes			
Sirota and Juanchich, 2011	Yes		Yes	
Siegrist and Keller, 2011	Yes/No [f]			
Hill and Brase, 2012	Yes			
Garcia-Retamero and Hoffrage, 2013	Yes			
Johnson and Tubau, 2013	Yes/No [g]			
Lesage et al., 2013			Yes	
Sirota et al., 2014a		Yes	Yes	Yes/No [b]

Note that variation exists between the specific context and numbers used across studies, as well as specific measures and criteria used to determine low vs. high performers (see text for additional details, and original articles for full problems and explanations).

**It is important to note that YES with normalized versions does not imply "good" reasoning, with most higher ability participants typically below 30% correct response.*

[I] *CRT, Cognitive Reflection Test (Frederick, 2005).*

[a] *YES with simple versions; NO with complex versions (floor effect).*

[b] *YES with REI (rational-experiential inventory; rational thinking); NO with CAOMTS (actively open-minded thinking).*

[c] *Information was normalized, but problems manipulated to require only simple single-step arithmetic.*

[d] *Higher numerate benefited more from causal manipulation used in Krynski and Tenenbaum (2007).*

[e] *NO with REI.*

[f] *YES in study 1; NO in study 2 (though clear trend).*

[g] *YES with complex text; NO with short, simple text.*

students tested in most studies, as well as by children (Zhu and Gigerenzer, 2006). At the same time, there is also evidence that many people either lack the cognitive clarity or are unwilling to invest the needed cognitive effort into even the simplest whole number arithmetic (addition, subtraction). For example, confirming their "mental steps hypothesis," Ayal and Beyth-Marom (2014) showed that performance drops sharply when more than a single numerical operation is required, even if these operations are little more than a series of simple additions. Related findings were observed in the "defective nested sets" study reported in Girotto and Gonzalez (2001, study 5) which presented a partitive structure [but with ($\neg H \& \neg D$) instead of ($\neg H \& D$)], but which required an additional subtraction to solve. Together, these findings demonstrate that natural frequency facilitations are not simply about the clarity of the presented data, but are also about how easily the specifically presented components allow reasoners to generate the Bayesian solution (see also Barbey and Sloman, 2007).

Verbal Formulation and Irrelevant Information

As with normalized versions, manipulating the verbal context of a problem to align with existing world knowledge can improve performance. For example, Siegrist and Keller (2011, study 4; see also Sirota et al., 2014a, study 2; Chapman and Liu, 2009) showed that a less educated group from the general population was more than twice (13 vs. 26%) as likely to solve a "social" problem (people lie, have red nose) vs. a "medical" problem (have cancer, test positive). They suggested this group may focus on specific task information in a real-world context, and might have assumed they did not know enough about cancer or medical tests to solve the problem. There is also evidence that performance, especially by lower numerate reasoners, is impaired by the presence of unnecessarily descriptive words in the text (Johnson and Tubau, 2013). Other verbal manipulations, such as clarifying the meaning of "false positive," have also been suggested to improve performance (Cosmides and Tooby, 1996; Sloman et al., 2003; see also Fox and Levav, 2004). Sloman et al. (2003) found that irrelevant numbers in the problem did not impair performance with transparent frequency problems, and suggested that a frequency format "makes it easier for people to distinguish relevant from irrelevant ratios" (p. 304). However, a very frequently reported error with natural frequencies is that reasoners use the superordinate value of the problem or the new reference class presented in the question as the denominator

in their response (e.g., "100" in **Figure 1**; see Gigerenzer and Hoffrage, 1995; Macchi and Mosconi, 1998; Mellers and McGraw, 1999; Evans et al., 2000; Girotto and Gonzalez, 2001; Brase et al., 2006; Zhu and Gigerenzer, 2006), suggesting that irrelevant numbers may indeed bias responses on simple natural frequency problems.

Visual Representations

Sloman et al. (2003) also found that Euler circles did not further enhance performance with their frequency problem, and suggested that visuals only facilitate if nested-set relations are not already clear (see also Cosmides and Tooby, 1996). In contrast, Yamagishi (2003) found improvements on a natural frequency gemstone problem with a roulette-wheel diagram. Brase (2009) did not find a benefit of a Venn diagram with chance versions in a natural frequency structure, however an icon display did provide an additional benefit beyond the frequency format (see also Brase, 2014). Complementary results by Garcia-Retamero and Hoffrage (2013) also showed benefits of visual aids above and beyond the use of natural frequencies. Garcia-Retamero et al. (2015) further showed that visual aids are particularly beneficial to lower numerate reasoners, and may also improve their metacognitive judgment calibration. Contrasting with the above, Sirota et al. (2014b) failed to find a benefit with several types of visuals. In brief, while some facilitation with visual aids has been reported with natural frequencies, current evidence is conflicting and suggests that other factors are likely interacting with the effectiveness of these aids.

Chances

Although initial reports implied that naturally sampled chances were as easily represented as naturally sampled frequencies (Girotto and Gonzalez, 2001), more recent studies show that this might not be the case (Brase, 2008, 2014; Sirota et al., 2015a). This would be in line with more general literature on the difficulties that people have learning and understanding probabilities (e.g., Garfield and Ahlgren, 1988; Gigerenzer et al., 2005; Morsanyi et al., 2009; Morsanyi and Szücs, 2015). It would also imply that the lack of difference between probability and proportion formulations with normalized data (see "*Number of Events*" above in Section Reasoning with Normalized Formats) is not because these formats are equally well (or poorly) understood, but rather that the difference is being masked by another more fundamental difficulty with normalized information (carrying out fraction arithmetic; understanding or identifying the requested relations). Interestingly, participants who interpret naturally sampled "chances" as frequencies outperform those individuals who interpret them as single-event probabilities (Brase, 2008, 2014). Also of interest, more recent evidence suggests the relevant "interpretation" may be at the problem level (in terms of set relations) rather than at the format level (in terms of frequencies) (Sirota et al., 2015a).

Individual Differences

It has been argued that the wide variability reported with natural frequency problems can be attributed to individual differences in ability or motivation (Brase et al., 2006; Barbey and Sloman, 2007). In line with this suggestion (and summarized in **Table 2**), better performance with natural frequencies has been observed by individuals higher in *cognitive reflection* (Sirota and Juanchich, 2011; Lesage et al., 2013; Sirota et al., 2014a; measured with the CRT); *fluid intelligence* (Sirota et al., 2014a; measured with Raven's matrices); preference for *rational thinking* (Sirota et al., 2014a; measured with the REI), *education level* (Brase et al., 2006; Siegrist and Keller, 2011; though see Hoffrage et al., 2015), and *numeracy* (Chapman and Liu, 2009; Sirota and Juanchich, 2011; Hill and Brase, 2012; Garcia-Retamero and Hoffrage, 2013; Johnson and Tubau, 2013; Garcia-Retamero et al., 2015; McNair and Feeney, 2015). These higher ability individuals often perform quite well, although the success of even these more capable individuals varies widely across studies.

While some of the between-study variation with natural frequencies can be captured by these individual differences factors, the strong relations observed with "higher ability" reasoners also raises some questions. Why are general intelligence, cognitive reflection, and numeracy so consistently relevant on such an arithmetically simple task, especially one in which the "structural transparency" of the task is such a well-toted facilitator? Indeed, due to the base-rate preservation, there is no need for a fully fleshed out representation of the entire problem structure, and attention need only be allocated to two pieces of information, ($H\&D$) and ($\neg H\&D$). Together, performance on these "simple" problems implies that, beyond simple text processing and whole-number arithmetic, there may be a particular logical difficulty inherent in these problems that is often overlooked.

Common Processing Demands: Quantitative Backward Reasoning

Early studies of Bayesian inference with the medical diagnosis task were specifically directed at understanding how individuals (e.g., physicians) diagnosis disease given a prior distribution and an imperfect predictor (a test result) (Casscells et al., 1978; Eddy, 1982). More recently, logical and set operations have been identified as a useful strategy for performing these Bayesian inferences (e.g., Sloman et al., 2003), however, we believe that the particular nature of the required set operations has been underemphasized in recent studies. More specifically, we suggest that a particularly difficult stage of Bayesian problem solving is performing a *backward* (diagnostic) inference (van den Broek, 1990; Oberauer and Wilhelm, 2000; Lagnado et al., 2005; Fernbach et al., 2011; Sloman and Lagnado, 2015)—in the medical diagnosis problem, working backward from a positive test result (effect) to the likelihood of being infected (cause), when information is provided in the forward cause → effect direction. For example, querying the model in **Figure 2B** in the direction opposite from which it was formed (i.e., infected → test positive), implies a change in the specific role (focal subset or reference class) of previously associated categories (or a change of *focus*; Dubois and Prade, 1997; Baratgin and Politzer, 2010), or a change in the direction of the causal link (test positive → infected).

This process can be facilitated with questions which guide reasoners through the search and selection process, for example, with integer pair question formats which prompt the reasoner

for two separate numbers rather than a single percentage, and perhaps even more so if the reference class is prompted prior to the focal subset (Girotto and Gonzalez, 2001). It is interesting to note that with natural frequencies this symmetrical confusion is reduced for the first term of the integer pair ("6 out of 24"): "*Among infected, 6 test positive*" to "*Among positive, 6 are infected,*" though the more challenging asymmetrical inference still remains for the reference class. This is consistent with the suggestion that one of the biggest challenges may be getting reasoners focused on the correct reference set of positive testers (Evans et al., 2000; Girotto and Gonzalez, 2001). While this particular logical difficulty (backward reasoning or quantification of backward relations) has not been directly demonstrated in Bayesian word problems, similar explanations have been used to successfully account for performance in other reasoning tasks (Evans, 1993; Barrouillet et al., 2000; Oberauer and Wilhelm, 2000; Oberauer et al., 2005; Oberauer, 2006; Waldmann et al., 2006; Sloman and Lagnado, 2015), suggesting it may also be a key stumbling block in Bayesian reasoning. This explanation might also help to explain why reasoning can be improved with manipulations which encourage the experience of a scenario from multiple perspectives, such as "interactivity" in other Bayesian tasks (Vallée-Tourangeau et al., 2015a) and the "perspective effect" in the Monty Hall dilemma (for review see Tubau et al., 2015), which would help to facilitate the backward inference. This suggestion is also in line with results of the "short menu" natural frequencies reported in Gigerenzer and Hoffrage (1995), which directly presented both ($H\&D$) and (D) in the problem, thereby eliminating all arithmetic, but which led to a negligible benefit compared with "standard menu" natural frequencies which require ($H\&D$) + ($\neg H\&D$) to compute (D).

Summary

Taken as a whole, evidence reviewed above is consistent with the claim that normative responding is generally improved by facilitating comprehension of both presented and requested information (e.g., presenting what is needed, removing what is irrelevant, using questions which guide the reasoner) and, relatedly, minimizing the number of explicit cognitive (logical and numerical) operations required to move from problem to solution. Likewise, increasing the cognitive capacity, relevant skills, or effort of the reasoner, will generally lead to more Bayesian responses. Individuals who are more drawn toward quantitative, analytical thinking are more likely to solve these problems. This summary is consistent with a general nested-sets hypothesis, which states that any manipulation which facilitates the representation of relevant set information will generally enhance performance (Barbey and Sloman, 2007). At the same time, simply representing the relevant *qualitative* relations amongst nested sets will not get you a Bayesian response. These relations must also be accurately *quantified*, along with the correctly identified backward (posterior) inference.

More generally, understanding *why* these facilitations work as they do requires consideration of the processes in which reasoners are engaged. Ultimately, a reasoner needs to provide the requested ratio in the requested form, but arriving at this point requires the successful completion of a series of intermediate subtasks. We believe that a better understanding of successful, or failed, Bayesian problem solving can be obtained by considering: (1) How a nested-set "structure" comes to be represented by a reasoner (whether transparently presented in the problem or not); (2) What additional computational requirements are required once the structure of the problem is made "transparent" (or transparently represented by a reasoner); and (3) Who is more likely to be driven toward and successfully operate over this quantified, abstract level of reasoning. In the next section we outline one suggestion of how to conceptualize these processing requirements.

Bayesian Problem Solving, from Comprehension to Solution

Solving a Bayesian word problem is a process, from the presented words and numbers, through the representations and computations invoked to transform presented information into requested ratios. As we outline below, we suggest that this process can be productively understood, at least in part, from the perspective of mathematical problem solving (for reviews see Kintsch and Greeno, 1985; Schoenfeld, 1985; LeBlanc and Weber-Russell, 1996; Mayer, 2003). On this view, the task is conceived as two interrelated processes: Text comprehension and problem solving. More specifically, successful reasoning depends essentially on comprehending the presented and requested information, and more importantly the *relation* between the two (i.e., the space between what is provided and requested). This comprehension then drives any logical or numerical computations necessary in order to reduce this space, ending with a final numerical response. A basic framework for understanding this process is outlined in **Figure 3**.

Two basic assumptions of this framework are (1) while the input may be the same, the (levels of) representations that reasoners operate over differ, and (2) specific individual skills or capacities and specific problem formulations can trade spaces. That is, the probability of successfully solving the problem will depend on the complexity of the provided information, as well as the number and complexity of the steps required to close this gap. Crucially, this process, and its relative difficulty, also depends on the abilities, tendencies, and skills of the problem solver (Schoenfeld, 1985, 1992; Cornoldi, 1997; Mayer, 1998, 2003; Swanson and Sachse-Lee, 2001; Passolungh and Pazzaglia, 2004). Succinctly, what is difficult for one person may not be difficult for another. In the remainder of this section, we more fully explicate this task-individual interaction as it unfolds during the reasoning process, from comprehension and computation to solution.

Comprehension of Presented and Requested Information

Solving a written Bayesian inference problem begins with text *comprehension*. Working memory serves as a buffer where recently read propositions and information activated in long-term memory are integrated into the internal model under construction (e.g., Just and Carpenter, 1992; Carpenter et al., 1995; Ericsson and Kintsch, 1995; Daneman and Merikle, 1996;

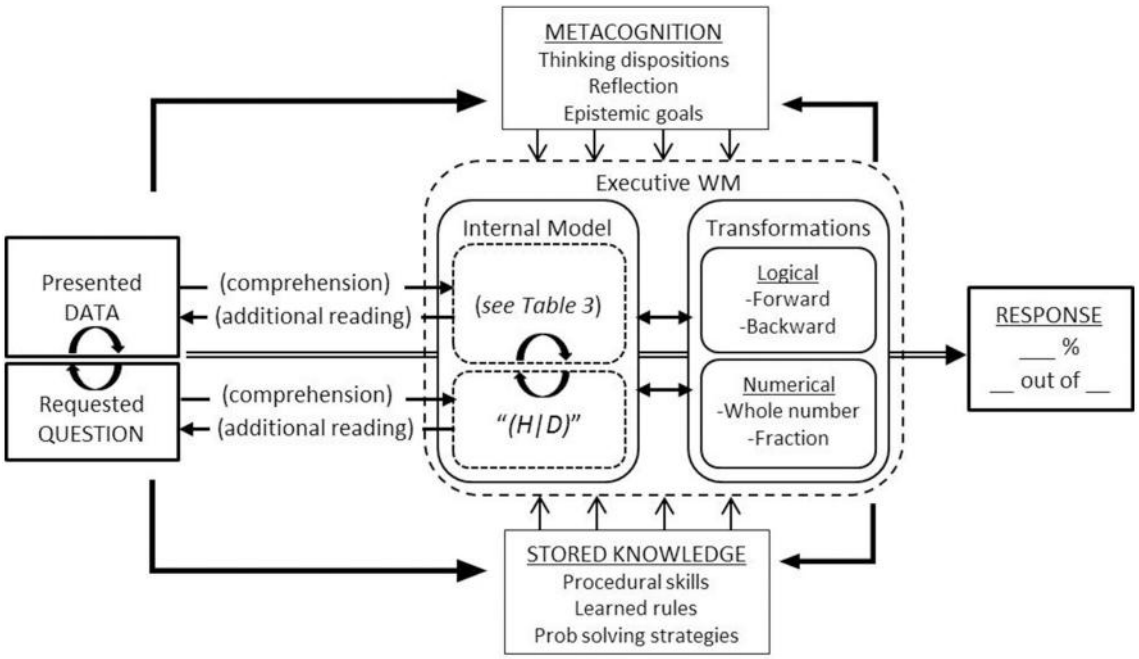

FIGURE 3 | Framework for understanding Bayesian word problem solving. The task (left) is conceived as the presented data and the requested question. A text comprehension process gives rise to an initial internal model of the data (including inferences not directly in the text; see **Table 3**). The comprehension of the question "*(H|D)*" initiates a goal-oriented search (though internal representation and problem text) for the requested relations, along with logical and numerical computations aimed at deriving information not directly available in the text. The processing of the task also activates metacognitive dispositions as well as potentially relevant stored knowledge and skills, both of which can influence both *what* information is processed and *how* that information is processed. A continuously updated working memory also provides reciprocal feedback to metacognition and calls for additional stored knowledge if needed. The final response given will depend on a complex interaction of the task formulation, metacognition, available knowledge and skills, and the efficiency of an executive working memory. For example, metacognition can influence the effort invested into the task, while stored knowledge can influence the relative effort required for a given individual.

Cain et al., 2001, 2004; Tronsky and Royer, 2003). Many inferences are automatically generated as a reader processes the symbolic words and numbers in the text (**Table 3**), resulting in an internal representation of the problem which may contain specific propositions included in the text itself, along with possible semantic, episodic, spatial, causal, categorical, and quantitative inferences, any of which can serve as the basis for upstream reasoning (e.g., Nesher and Teubal, 1975; van Dijk and Kintsch, 1983; Kintsch and Greeno, 1985; Murray et al., 1993; Graesser et al., 1994; LeBlanc and Weber-Russell, 1996; Vinner, 1997; Reyna et al., 2003; Reyna and Brainerd, 2008; Thompson, 2013). These levels of representations are activated to varying degrees, and may be either implicit or explicit (or not present at all) within a reasoner's model of the problem (cf. Johnson-Laird, 1983). As we return to below, given the multiple levels of information that *can* be represented, one challenge is getting a limited attention focused on the most relevant information for problem solving.

It is with the reading of the question that the relevance of any initially represented problem information becomes apparent. The formulation of the question therefore plays a crucial role in Bayesian problem solving (Schwartz et al., 1991; Macchi, 1995; Girotto and Gonzalez, 2001). In general, the question provides two specific prompts: (a) verbal cues corresponding to the required categorical relations, or ratio, to be provided, and (b) the format in which the quantified response should be provided. For example, a typical natural frequency question prompts the reasoner to find two whole numbers "__ *out of* __" corresponding to "*among infected, how many positive*"; while a standard probability question demands a single percentage "__%" corresponding to "*infected if positive*"). This question comprehension triggers a goal-oriented search (through memory representations and the problem text) for the specific relations requested, along with more directed inferences and arithmetical computations targeted at deriving information not directly provided in the text (see Section Logical and Numerical Computations).

Regardless of the problem format, we expect that with relatively little effort most literate reasoners comprehend the basic situation (some people take a test for a disease), form simple categorical-numerical associations (inf[10%]; inf-pos[60%]), and make some simple forward inferences (**Table 3**). Ultimately, however, providing a precise Bayesian response requires accurate representation and quantification of appropriate set-subset relations (i.e., *H&D, D*), irrespective of the problem content. Comprehension of the *categorical* subset structure can be facilitated by presenting natural frequencies, highlighting causal structure, removing irrelevant information, providing visual

TABLE 3 | Examples of inferences and levels of encoding generated while reading a Bayesian word problem.

		Hypothetical knowledge and inferences	
		Normalized	**Natural frequencies**
Prior knowledge or beliefs	→	Infections cause positive tests. Medical tests are usually accurate. A positive test should indicate infection. …	
Forward categorical	→	Some people are infected. Some of the infected test positive. Some of the infected do not test positive. Some of the not infected also test positive. …	
Backward categorical	→	Some of the positives are infected. Some of the positives are not infected. A positive test does not necessarily mean infected. …	
Non-integrated categorical-numerical association	→	Infected [10%] Infected-Positive [60%] Not infected [90%] Not infected-Positive [20%] …	Total [100] Infected [10] Infected-Positive [6] Not infected [90] Not infected-Positive [18] …
Forward quantitative	→	60% of 10% = 6% are both inf and pos 20% of 90% = 18% are not-inf and pos 6% + 18% = 24% of people are pos …	6 people are both inf and pos 18 people are both not-inf and pos 6 + 18 = 24 people are pos …
Backward quantitative	→	Of 24% pos, 6% are infected Of 24% pos, 18% are not infected If pos, chances of inf are 25% (6/24%) …	Of 24 positive, 6 are infected Of 24 positive, 18 are not infected If pos, chances of inf are 6 of 24 …

Inferences may be spontaneously generated during text comprehension, or prompted as a result of the question, and may be either implicit or explicit (or not present at all) within a reasoner's model of the problem. A variety of biased responses are possible based on erroneous or irrelevant prior knowledge or beliefs, non-integrated representations, or attention to inappropriate levels of information. Inf, infected; Pos, positive test.

diagrams, asking questions which direct attention toward relevant information, etc. Accurate comprehension of the *quantified* values (strength) of these relations is facilitated when numerical information is presented with a natural sampling structure. In the case of relative frequencies (or non-partitive probability information), on the other hand, correct *qualitative* representation of the subset structure may coincide with incorrect or incomplete *quantification* of these relations. For example, in **Figure 1** it is feasible that reasoners understand the 60% hit rate to be a subset of the 10% infected, but the precise comprehension of this value requires more demanding, rule-based transformations (although some higher numerate reasoners may rather automatically perform "simple" computations such as 60 of 10% = 6%).

Individuals with more cognitive capacity will tend to more deeply processes the text (defined by the number of accurate and successfully integrated inferences; for reviews see van Dijk and Kintsch, 1983; Graesser et al., 1994), and likewise end up with a more thorough representation of the available information (both its content and structure) and the task goal as comprehended from the question. Accordingly, higher cognitive capacity or higher cognitive reflection will facilitate comprehension of a Bayesian task, at least to some extent (see **Table 2**). We also suggest that the processing of the task gives rise to a metacognitive assessment reflecting motivation and confidence that the problem can be solved ("*can I do this?*," "*do I want to do this?*"), which will help to guide subsequent problem solving behavior (e.g., Schoenfeld, 1992; Cornoldi, 1997; Mayer, 1998; Thompson, 2009; also Garcia-Retamero et al., 2015).

At the same time, information from long-term memory is being integrated into working memory—including prior knowledge of causal relationships (Krynski and Tenenbaum, 2007), situational familiarity (Siegrist and Keller, 2011), or other primed categories (Kahneman et al., 1982)—which leads to different levels at which a problem can be represented (**Table 3**), only some of which are relevant for solving the problem. Therefore, getting focused on the relevant set relations and their numerical values, while inhibiting ultimately irrelevant contextual details and prior beliefs (e.g., about the validity of medical tests), is crucial. The ability to do so should accordingly depend in part on executive functions and working memory (see Barrett et al., 2004; Evans and Stanovich, 2013). It is further known that engaging a Bayesian problem also triggers stored knowledge associated with problem solving strategies

and mathematical concepts and procedures, which act to bias attention to different levels of information within the task, for example, by leading the problem solver to analyze the text in a way which may differ from how they read stories or other news (e.g., Newell and Simon, 1972; Nesher and Teubal, 1975; Kintsch and Greeno, 1985; Anderson, 1993; Ericsson and Kintsch, 1995; Geary, 2006). In this vein, higher cognitive reflection and numeracy may also serve to bias attention toward relevant numerical information and away from irrelevant descriptive information, or more generally to relevant abstract formal relations amongst problem data rather than literal problem features (Spilich et al., 1979; Chi et al., 1981; Hegarty et al., 1995; Vinner, 1997; Peters et al., 2007; Dieckmann et al., 2009; Johnson and Tubau, 2013). This can help account for the consistent relationship between numeracy and Bayesian reasoning with natural frequencies, including interactions with non-numerical factors (**Table 2**).

Logical and Numerical Computations

Information which is needed but not directly provided must be derived. The transformations needed to produce this information can be numerical or logical. In standard Bayesian inference tasks, numerical computations typically include whole number and/or fraction arithmetic. Whole number arithmetic is a skill that tested populations (university undergraduates; medical professionals) can be assumed to possess. At the same time, it has been shown that, even with natural frequencies, performance drops quickly when more than a single whole number addition or subtraction is required (e.g., Girotto and Gonzalez, 2001; Ayal and Beyth-Marom, 2014). Curiously, if normalized data allows the posterior relation ($H|D$) to be derived with a single whole number subtraction, performance is actually quite high, even for less numerate reasoners (Ayal and Beyth-Marom, 2014). However, it is not clear if this latter finding is due to the reduced computational demands, or from the easier representation of how to derive the standard posterior relation (e.g., by eliminating the need to perform the backward inference). While it is often assumed that fraction arithmetic (e.g., multiplying two percentages) is a skill possessed by tested populations, this may not be the case (Paulos, 1988; Butterworth, 2007; Reyna and Brainerd, 2007, 2008), as some evidence suggests (e.g., Juslin et al., 2011; Ayal and Beyth-Marom, 2014). In brief, current evidence indicates that a single whole number addition adds minimal burden to the task; more than a single operation regardless of type greatly reduces performance; and it is not clear to what extent typically tested reasoners possess the procedural skills for carrying out single multiplicative integrations.

Required computations can also be logical. As previously identified, one crucial step for solving the posterior Bayesian question which may be particularly difficult is the backward inference (*test positive→ infection*), from the initially forward relations (*infection→ test positive* and *no-infection→ test positive*), or otherwise identifying the newly required reference class and focal subset (more likely prompted by the two-term integer pair question). The specific difficulties deriving and quantifying a diagnostic inference from predictive relations is well-known from causal reasoning tasks (e.g., van den Broek, 1990; Lagnado et al., 2005; Fernbach et al., 2011; Sloman and Lagnado, 2015). The asymmetry between the quantification of the relations presented and those requested requires reasoners to inhibit the precise quantifiers attached to the original relations and update the precise quantifier corresponding to the newly required relations (e.g., corresponding to the strength of the *test positive→ infection* relation). As mentioned, natural frequencies may alleviate part of this asymmetrical confusion (for the first term of the ratio, (*H&D*); i.e., "positive among infected" = "infected among positive"), but the more challenging identification of the reference class still remains. The ability to identify and quantify this new relation should accordingly be moderated by executive functions and skill in mathematical and logical reasoning.

Arriving at a Final Response

As outlined above, the final response provided by a reasoner reflects the confluence of a comprehension and problem solving process engaged by an individual with a particular set of skills and dispositions (**Figure 3**). The accuracy of this response will therefore depend on the level of problem comprehension and, relatedly, on the individual skills available to inspect and appropriately transform a dynamically updated internal model. A small set of rather systematic errors often account for a large proportion of erroneous responses, the most widely reported in either format being the direct selection of the hit rate, or the "inverse fallacy" (see Kahneman and Tversky, 1972; Koehler, 1996; Villejoubert and Mandel, 2002; Mandel, 2014a). Nevertheless, it is still not clear whether this results from errors understanding logical categorical relations (e.g., Wolfe, 1995; Villejoubert and Mandel, 2002) vs. superficial problem solving strategies (e.g., matching; see Evans, 1998; Stupple et al., 2013). That is, a variety of sources of failures—from erroneous comprehension of the particular relations requested to difficulties inhibiting irrelevant, previously primed information—could account for common errors, and it is not necessarily the case that the underlying cause is the same for all reasoners. The proposed framework might help to improve understanding of the causes of observed failures.

One of the main thrusts of the nested-sets hypothesis is that *if* the formulation of the problem triggers awareness of the set structural relations amongst the presented categories, *then* general cognitive resources can employ elementary set operations to mimic a Bayesian response. Musing on this possibility, Sloman et al. (2003, p. 307) suggested:

> "A question that might be more pertinent is whether our manipulations changed the task that participants assigned themselves. In particular, manipulations that facilitate performance may operate by replacing a non-extensional task interpretation, like evidence strength, with an extensional one (Hertwig and Gigerenzer, 1999). Note that such a construal of the effects we have shown just reframes the questions that our studies address: under what conditions are people's judgments extensional…"

In this sense, natural frequencies (or other nested-sets facilitations) might shift reasoners into a more analytical

mode of thinking due to the stronger match between presented information and the available reasoning tools of the participants, a mode we have suggested might be more automatically adopted by individuals with higher mathematical or cognitive skills. Considered from a problem solving perspective, one factor separating successful from unsuccessful reasoners may be the way they formulate and answer three crucial, interrelated questions: "*What information do I have available?*" (means), "*What information do I need to provide?*" (ends), and "*What steps do I need to take to close this gap?*" (solution plan). The relative difficulty answering these questions will of course depend on the complexity of the provided information along with the number and complexity of the required steps, and also on the individual capacities and skills of the reasoner.

More specifically, the present review suggests at least three crucial sources of difficulty for arriving at a correct Bayesian response: (1) accurately quantifying the relevant forward categorical relations of the problem, (2) accurately performing the needed backward inference, including identifying and quantifying the relevant reference class, and (3) formulating and executing and multi-step plan required for transforming presented data into the requested ratio. Each of these requirements are facilitated with natural frequencies, and become increasingly difficult with normalized data. As previously commented, performance in the latter case remains low even for participants higher in cognitive capacity or higher in numeracy, suggesting that success on these problems depends on specific skills not adequately acquired, or not spontaneously employed, by most of the participants in reviewed studies.

Hence, looked at from another direction, if the objective is to narrow the gap between human performance and Bayesian prescriptions when reasoning from explicit statistics, part of the remedy is to get participants to think more mathematically (see also Zukier and Pepitone, 1984; Schwartz et al., 1991). People are not born able to deal with abstract symbolic words and numbers. Both reading and math ability develop over time with education and practice. Even with extensive education, many individuals still fail to attain the relevant conceptual and procedural knowledge for dealing with ratios (Paulos, 1988; Brase, 2002; Ni and Zhou, 2005; Butterworth, 2007; Reyna and Brainerd, 2007, 2008; Siegler et al., 2011, 2013), a difficulty which is exacerbated when these number concepts are embedded in textual scenarios (e.g., Kirsch et al., 2002). Ultimately, therefore, deficits in explicit statistical reasoning may need to be addressed at the level of mathematics education. This remedy is not as immediate as simply reformulating a problem with natural frequencies, but in the long-term this may be a necessary way to obtain the levels of performance with which we can be satisfied.

Future Directions

The way to proceed toward a better understanding of probabilistic reasoning potentials and pitfalls depends on the specific question of interest. A variety of questions and paradigms have been addressed in this special issue on Bayesian reasoning ("Improving Bayesian Reasoning: What Works and Why?"), ranging from alternative probabilistic standards (Douven and Schupbach, 2015) to important real world issues (Navarrete et al., 2014). While many of these have focused on Bayesian word problems, other paradigms have also been discussed including "uncertain deduction" (Cruz et al., 2015; Evans et al., 2015), the Monty Hall Dilemma (Tubau et al., 2015), and the Sleeping Beauty problem (Mandel, 2014b) (for brief overviews see Mandel, 2014a; Girotto and Pighin, 2015; Juslin, 2015; McNair, 2015; Vallée-Tourangeau et al., 2015b). With respect to Bayesian word problems, many authors have expressed similar views regarding the problem-solving nature of these tasks (e.g., McNair, 2015; Sirota et al., 2015c; Vallée-Tourangeau et al., 2015b), which echo many themes presented in this review. In what follows, we offer suggestions on ways to progress in this later, problem-solving paradigm, but which may also be applicable to other paradigms as well.

Moving forward, we believe there is a need to shift perspective from the facilitators of Bayesian reasoning to more process-oriented measures aimed at uncovering the strategies evoked by successful and unsuccessful reasoners, and the stages in the problem solving process at which these differences emerge (for one proposal see De Neys and Bonnefon, 2013). To this general end, we suggest that tools from the mathematical problem solving approach might be productively applied to research on Bayesian reasoning. For example, the "moving window" (Just et al., 1982; see also De Neys and Glumicic, 2008) and online recognition paradigms (Thevenot et al., 2004; Thevenot and Oakhill, 2006) can be used to assess comprehension at different stages of problem solving, as well as *when* calculations are made throughout the reasoning process. These methods, which control or limit access to specific pieces of information, can be applied to help determine the relative difficulty of representing vs. quantifying relevant structural relations, both forward and backward.

Other process methods such as eye-tracking and recall tests are also frequently used to measure how attention is allocated during the problem solving process, which can be used to gauge the weight a reasoner assigns to different pieces of information in the text (Mayer, 1982; Hegarty et al., 1992, 1995; Verschaffel et al., 1992; for overviews see Mayer et al., 1992; LeBlanc and Weber-Russell, 1996; also De Neys, 2012). The use of protocol and error analyses have also proven effective in studies of mathematic problem solving and other areas of decision making and reasoning (e.g., Kuipers and Kassirer, 1984; Chi, 1997; Arocha et al., 2005; De Neys and Glumicic, 2008; Kingsdorf and Krawec, 2014), but apart from some notable exceptions (e.g., Gigerenzer and Hoffrage, 1995; Zhu and Gigerenzer, 2006) have thus far played only a limited role in Bayesian reasoning research. These methods can offer substantial insight into the level of information and cognitive processes that successful vs. unsuccessful reasoners engage (see also McNair, 2015). These tools can also be adapted to address questions about how participants are interpreting the tasks given to them, and the extent to which they are attempting to produce precisely computed responses vs. numerical estimates based on future uncertainties.

Finally, we suggest that these approaches be adopted alongside a strong commitment to individual differences (e.g., Stanovich et al., 2011; Del Missier et al., 2012; De Neys and Bonnefon,

2013; Klaczynski, 2014). More specifically, processing measures should look not only to establish where in the reasoning process correct and incorrect solvers depart, but also as a function of specific individual differences in ability, disposition, and requisite skills. Methods from mathematical problem solving could help to confirm or clarify existing proposals for the relevance and relative influence of these individual differences (De Neys and Bonnefon, 2013; Klaczynski, 2014).

Conclusion

Successive waves of Bayesian reasoning research have gradually revealed that non-Bayesian responses in statistical and probabilistic word problems arise not out of biased heuristics guiding belief revision, but rather out of failed analytical processing operating over specific task structures. Even the simplest Bayesian word problems are not solved automatically, but rather involve deliberate analytical processing of the verbal and numerical structure of the task, and the subsequent logical and numerical transformations of presented data into requested relations. The formulation of the task can influence the specific types and number of inferences required to solve the problem. Hence, reducing the distance between problem and solution (mental steps) and, independently, making clear what is relevant for problem solving will generally facilitate performance. At the same time, individual differences will moderate the effect of these computational demands, as the effort required is relative to the availability of cognitive resources and relevant stored knowledge. That is, reducing processing demands and/or increasing processing resources are two complimentary means to the same end—a Bayesian response.

We have argued that a better understanding of this task-individual pair can be gained by shifting attention to the processing requirements needed to compute the Bayesian response, and the processing strategies which may be adopted by different reasoners. The proposed account, borrowed from the mathematical problem solving literature, suggests that this begins with text comprehension, an inferential and integrative process which draws on cognitive capacity and previous knowledge and skills. This gives rise to an initial problem representation, along with metacognitive assessments, which serves as the basis for the subsequent question comprehension and problem-solving behavior. We have also identified two crucial factors in this process which are likely to cause particular difficulty for many reasoners: (1) accurately *quantifying* the relevant structural relations amongst hierarchically embedded subset categories, and (2) quantifying the *backward inference* mandated by the asymmetrical direction of presented (infection→test positive) and requested (test positive→infection) information. Accordingly, interventions targeting these factors are likely to have the greatest success (i.e., natural frequencies, familiar causal relationships, guided questions), and task-relevant individual skills and abilities (numeracy, logical capacity, disposition toward analytical thinking) are likely to interact with the effectiveness of these interventions.

Given the multiple representations that Bayesian problems afford—spatially as nested sets, numerically as proportions, formally in Bayes theorem—they offer a natural link to theories of reasoning with proportional information. Accordingly, we have suggested that understanding why individuals succeed or fail on these problems can be partially anchored in the field of mathematical cognition, which has long emphasized the difficulties in learning and using ratio information, along with the importance of metacognition and executive working memory for successfully integrating different set-subset relations and for dealing with numerical information in varying contexts and formats. We believe that this complimentary perspective, and the tools it employs, can help guide a more process-oriented approach aimed at more precisely understanding where reasoning with explicit categorical and numerical information goes astray, and how the individual reasoner can be redirected to align with Bayesian norms.

Acknowledgments

The authors would like to thank Ulrich Hoffrage and Jean Baratgin for their very helpful and constructive comments on previous drafts of this paper. This work was supported by grants from the Generalitat de Catalunya (FI-DGR 2011) awarded to the first author, the Spanish Ministry of Economics and Competitiveness (PSI2012-35703) and (PSI2013-41568-P), and 2014 SGR-79 from the Catalan Government.

References

Anderson, J. R. (1993). Problem solving and learning. *Am. Psychol.* 48, 35–44. doi: 10.1037/0003-066X.48.1.35

Arocha, J. F., Wang, D., and Patel, V. L. (2005). Identifying reasoning strategies in medical decision making: a methodological guide. *J. Biomed. Inform.* 38, 154–171. doi: 10.1016/j.jbi.2005.02.001

Ayal, S., and Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgm. Decis. Mak.* 9, 226–242.

Baratgin, J. (2002). Is the human mind definitely not bayesian ? A review of the various arguments. *Curr. Psychol. Cogn.* 21, 653–682.

Baratgin, J., and Politzer, G. (2006). Is the mind Bayesian? The case for agnosticism. *Mind Soc.* 5, 1–38. doi: 10.1007/s11299-006-0007-1

Baratgin, J., and Politzer, G. (2010). Updating : a psychologically basic situation of probability revision. *Think. Reason.* 16, 253–287. doi: 10.1080/13546783.2010.519564

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–297. doi: 10.1017/S0140525X07001653

Bar-Hillel, M. (1983). "The base-rate fallacy controversy," in *Decision Making Under Uncertainty,* ed R. W. Scholz (Amsterdam: Elsevier), 39–61.

Barrett, L. F., Tugade, M. M., and Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychol. Bull.* 130, 553–573. doi: 10.1037/0033-2909.130.4.553

Barrouillet, P., Grosset, N., and Lecas, J. F. (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition* 75, 237–266. doi: 10.1016/S0010-0277(00)00066-4

Barton, A., Mousavi, S., and Stevens, J. R. (2007). A statistical taxonomy and another "chance" for natural frequencies. *Behav. Brain Sci.* 30, 255–256. doi: 10.1017/s0140525x07001665

Bayes, T. (1764). An essay toward solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond.* 53, 370–418. doi: 10.1093/biomet/45.3-4.296

Brase, G. L. (2002). "Bugs" built into the system: how privileged representations influence mathematical reasoning across lifespan. *Learn. Individ. Dif.* 12, 391–409. doi: 10.1016/S1041-6080(02)00048-1

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychon. Bull. Rev.* 15, 284–289. doi: 10.3758/PBR.15.2.284

Brase, G. L. (2009). Pictorial representations and numerical representations in Bayesian reasoning. *Appl. Cogn. Psychol.* 23, 369–381. doi: 10.1002/acp.1460

Brase, G. L. (2014). The power of representation and interpretation: doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *J. Cogn. Psychol.* 26, 81–97. doi: 10.1080/20445911.2013.861840

Brase, G. L., Cosmides, L., and Tooby, J. (1998). Individuals, counting, and statistical inference: the role of frequency and whole-object representations in judgment under uncertainty. *J. Exp. Psychol. Gen.* 127, 3–21. doi: 10.1037/0096-3445.127.1.3

Brase, G. L., Fiddick, L., and Harries, C. (2006). Participant recruitment methods and statistical reasoning performance. *Q. J. Exp. Psychol.* 59, 965–976. doi: 10.1080/02724980543000132

Brase, G. L., and Hill, W. T. (2015). Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Front. Psychol.* 6:340. doi: 10.3389/fpsyg.2015.00340

Butterworth, B. (2007). Why frequencies are natural. *Behav. Brain Sci.* 30, 259. doi: 10.1017/S0140525X07001707

Cain, K., Oakhill, J. V., Barnes, M. A., and Bryant, P. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Mem. Cognit.* 29, 850–859. doi: 10.3758/BF03196414

Cain, K., Oakhill, J. V., and Bryant, P. (2004). Children's reading comprehension ability: concurrent prediction by working memory, verbal ability, and component skills. *J. Educ. Psychol.* 96, 31–42. doi: 10.1037/0022-0663.96.1.31

Carpenter, P. A., Miyake, A., and Just, M. A. (1995). Language comprehension: sentence and discourse processing. *Annu. Rev. Psychol.* 46, 91–120. doi: 10.1146/annurev.ps.46.020195.000515

Casscells, W., Schoenberger, A., and Graboys, T. (1978). Interpretation by physicians of clinical laboratory results. *N. Eng. J. Med.* 299, 999–1000. doi: 10.1056/NEJM197811022991808

Chapman, G. B., and Liu, J. J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgm. Decis. Mak.* 4, 34–40.

Chater, N., and Oaksford, M. (2008). *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. New York, NY: Oxford University Press.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: a practical guide. *J. Learn. Sci.* 6, 271–315. doi: 10.1207/s15327809jls0603_1

Chi, M. T. H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cogn. Sci.* 5, 121–152. doi: 10.1207/s15516709cog0502_2

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., and Garcia-Retamero, R. (2012). Measuring risk literacy: the Berlin numeracy test. *Judgm. Decis. Mak.* 7, 25–47.

Cornoldi, D. L. C. (1997). Mathematics and metacognition: what is the nature of the relationship? *Math. Cogn.* 3, 121–139. doi: 10.1080/135467997387443

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

Cruz, N., Baratgin, J., Oaksford, M., and Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Front. Psychol.* 6:192. doi: 10.3389/fpsyg.2015.00192

Daneman, M., and Merikle, P. M. (1996). Working memory and language comprehension: a meta-analysis. *Psychon. Bull. Rev.* 3, 422–433. doi: 10.3758/BF03214546

Del Missier, F., Mantyla, T., and Bruine de Bruin, W. (2012). Executive functions in decision making: an individual differences approach. *Think. Reason.* 16, 69–97. doi: 10.1080/13546781003630117

De Neys, W. (2012). Bias and conflict : a case for logical intuitions. *Perspect. Psychol. Sci.* 7, 28. doi: 10.1177/1745691611429354

De Neys, W., and Bonnefon, J. F. (2013). The whys and whens of individual differences in thinking biases. *Trends Cogn. Sci.* 17, 172–178. doi: 10.1016/j.tics.2013.02.001

De Neys, W., and Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition* 106, 1248–1299. doi: 10.1016/j.cognition.2007.06.002

Dieckmann, N., Slovic, P., and Peters, E. M. (2009). The use of narrative evidence and explicit likelihood by decision makers varying in numeracy. *Risk Anal.* 29, 1473–1488. doi: 10.1111/j.1539-6924.2009.01279.x

Douven, I., and Schupbach, J. N. (2015). Probabilistic alternatives to Bayesianism: the case of explanationism. *Front. Psychol.* 6:459. doi: 10.3389/fpsyg.2015.00459

Dubois, D., and Prade, H. (1992). Evidence, knowledge, and belief functions. *Int. J. Approximate Reason.* 6, 295–319. doi: 10.1016/0888-613X(92)90027-W

Dubois, D., and Prade, H. (1997). "Focusing vs. belief revision: a fundamental distinction when dealing with generic knowledge," in *Qualitative and Quantitative Practical Reasoning of Lecture Notes in Computer Science* Vol. 1244, eds D. Gabbay, R. Kruse, A. Nonnengart, and H. Ohlbach (Berlin; Heidelberg: Springer), 96–107.

Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine: problems and opportunities," in *Judgment under Uncertainty: Heuristics and Biases,* eds D. Kahneman, P. Slovic, and A. Tversky (Cambridge: Cambridge University Press), 249–267. doi: 10.1017/CBO9780511809477.019

Ericsson, K. A., and Kintsch, W. (1995). Long-term working memory. *Psychol. Rev.* 102, 211–245. doi: 10.1037/0033-295X.102.2.211

Evans, J. S. (1993). The mental model theory of conditional reasoning: critical appraisal and revision. *Cognition* 48, 1–20. doi: 10.1016/0010-0277(93)90056-2

Evans, J. S. (1998). Matching bias in conditional reasoning: do we understand it after 25 years? *Think. Reason.* 4, 45–82. doi: 10.1080/135467898394247

Evans, J. S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* 59, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

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

Evans, J. S., and Stanovich, K. E. (2013). Dual process theories of cognition: advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685

Evans, J. S., Thompson, V. A., and Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Front. Psychol.* 6:398. doi: 10.3389/fpsyg.2015.00398

Fernbach, P. M., Darlow, A., and Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *J. Exp. Psychol. Gen.* 140, 168–185. doi: 10.1037/a0022100

Fox, C. R., and Levav, J. (2004). Partition–edit–count: naive extensional reasoning in judgment of conditional probability. *J. Exp. Psychol. Gen.* 133, 626–642. doi: 10.1037/0096-3445.133.4.626

Frederick, S. (2005). Cognitive reflection and decision making. *J. Econ. Perspect.* 19, 25–42. doi: 10.1257/089533005775196732

Garcia-Retamero, R., and Cokely, E. T. (2013). Communicating health risks with visual aids. *Curr. Dir. Psychol. Sci.* 22, 392–399. doi: 10.1177/0963721413491570

Garcia-Retamero, R., Cokely, E. T., and Hoffrage, U. (2015). Visual aids improve diagnostic inferences and metacognitive judgment calibration. *Front. Psychol.* 6:932. doi: 10.3389/fpsyg.2015.00932

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

Garfield, J., and Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: implications for research. *J. Res. Math. Educ.* 19, 44–63. doi: 10.2307/749110

Geary, D. C. (2006). "Development of mathematical understanding," in *Cognition, Perception, and Language*, Vol. 2, eds D. Kuhl and R. S. Siegler (New York, NY: John Wiley and Sons), 777–810.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond "heuristics and biases" in *European Review of Social Psychology*, Vol. 2, eds W. Stroebe and M. Hewstone (Chichester: Willey and Sons Ltd.), 83–115.

Gigerenzer, G., Hertwig, R., van den Broek, E., Fasolo, B., and Katsikopoulos, K. V. (2005). "A 30% Chance of Rain Tomorrow": how does the public understand probabilistic weather forecasts? *Risk Anal.* 25, 623–629. doi: 10.1111/j.1539-6924.2005.00608.x

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., and Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev.* 106, 425–430. doi: 10.1037/0033-295X.106.2.425

Gigerenzer, G., and Hoffrage, U. (2007). The role of representation in Bayesian reasoning: correcting common misconceptions. *Behav. Brain Sci.* 30, 264–267. doi: 10.1017/S0140525X07001756

Girotto, V., and Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition* 78, 247–276. doi: 10.1016/S0010-0277(00)00133-5

Girotto, V., and Gonzalez, M. (2002). Chances and frequencies in probabilistic reasoning: rejoinder to Hoffrage, Gigerenzer, Krauss, and Martignon. *Cognition* 84, 353–359. doi: 10.1016/S0010-0277(02)00051-3

Girotto, V., and Gonzalez, M. (2007). How to elicit sound probabilistic reasoning: beyond word problems. *Behav. Brain Sci.* 30, 268. doi: 10.1017/S0140525X07001768

Girotto, V., and Pighin, S. (2015). Basic understanding of posterior probability. *Front. Psychol.* 6:680. doi: 10.3389/fpsyg.2015.00680

Graesser, A. C., Singer, M., and Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychol. Rev.* 101, 371–395. doi: 10.1037/0033-295X.101.3.371

Hafenbrädl, S., and Hoffrage, U. (2015). Towards an ecological analysis of Bayesian inferences: how task characteristics influence responses. *Front. Psychol.* 6:939. doi: 10.3389/fpsyg.2015.00939

Hattori, M., and Nishida, Y. (2009). Why does the base rate appear to be ignored? The equiprobability hypothesis. *Psychon. Bull. Rev.* 16, 1065–1070. doi: 10.3758/PBR.16.6.1065

Hegarty, M., Mayer, R. E., and Green, C. E. (1992). Comprehension of arithmetic word problems: evidence from students' eye fixations. *J. Educ. Psychol.* 84, 76–84. doi: 10.1037/0022-0663.84.1.76

Hegarty, M., Mayer, R. E., and Monk, C. A. (1995). Comprehension of arithmetic word problems: a comparison of successful and unsuccessful problem solvers. *J. Educ. Psychol.* 87, 18–32. doi: 10.1037/0022-0663.87.1.18

Hertwig, R., and Gigerenzer, G. (1999). The "conjunction fallacy" revisited: how intelligent inferences look like reasoning errors. *J. Behav. Decis. Mak.* 12, 275–305.

Hill, W. T., and Brase, G. L. (2012). When and for whom do frequencies facilitate performance ? On the role of numerical literacy. *Q. J. Exp. Psychol.* 65, 2343–2368. doi: 10.1080/17470218.2012.687004

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition* 84, 343–352. doi: 10.1016/S0010-0277(02)00050-1

Hoffrage, U., Hafenbrädl, S., and Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Front. Psychol.* 6:642. doi: 10.3389/fpsyg.2015.00642

Johnson, E. D., and Tubau, E. (2013). Words, numbers, and numeracy: diminishing individual differences in Bayesian reasoning. *Learn. Ind. Diff.* 28, 34–40. doi: 10.1016/j.lindif.2013.09.004

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge: Cambridge University Press.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., and Caverni, J. P. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychol. Rev.* 106, 62–88. doi: 10.1037/0033-295X.106.1.62

Juslin, P. (2015). Controlled information integration and Bayesian inference. *Front. Psychol.* 6:70. doi: 10.3389/fpsyg.2015.00070

Juslin, P., Nilsson, H., Winman, A., and Lindskog, M. (2011). Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats. *Cognition* 120, 248–267. doi: 10.1016/j.cognition.2011.05.004

Just, M. A., and Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.* 99, 122–149. doi: 10.1037/0033-295X.99.1.122

Just, M. A., Carpenter, P. A., and Wooley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238. doi: 10.1037/0096-3445.111.2.228

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* 58, 697–720. doi: 10.1037/0003-066X.58.9.697

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. New York and Cambridge: Cambridge University Press.

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Kingsdorf, S., and Krawec, J. (2014). Error analysis of mathematical word problem solving across students with and without learning disabilities. *Learn. Disabil. Res. Pract.* 29, 66–74. doi: 10.1111/ldrp.12029

Kintsch, W., and Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychol. Rev.* 92, 109–129. doi: 10.1037/0033-295X.92.1.109

Kirsch, I. S., Jungeblut, A., Jenkins, L., and Kolstad, A. (2002). *Adult literacy in America: A first look at the findings of the National Adult Literacy Survey (NCES)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Klaczynski, P. A. (2014). Heuristics and biases: interactions among numeracy, ability, and reflectiveness predict normative responding. *Front. Psychol.* 5:665. doi: 10.3389/fpsyg.2014.00665

Kleiter, G. D. (1994). "Natural sampling: Rationality without base rates," in *Contributions to mathematical psychology, psychometrics, and methodology*, eds G. H. Fischer and D. Laming (New York, NY: Springer), 375–388.

Koehler, J. J. (1996). The base rate fallacy reconsidered: descriptive, normative, and methodological challenges. *Behav. Brain Sci.* 19, 1–53. doi: 10.1017/S0140525X00041157

Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability*. Oxford: Chelsea Publishing Co.

Krynski, T. R., and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *J. Exp. Psychol. Gen.* 136, 430–450. doi: 10.1037/0096-3445.136.3.430

Kuipers, B., and Kassirer, J. P. (1984). Causal reasoning in medicine: an analysis of a protocol. *Cogn. Sci.* 8, 363–385. doi: 10.1207/s15516709cog0804_3

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., and Sloman, S. A. (2005). "Beyond covariation: cues to causal structure," in *Causal Learning: Psychology, Philosophy and Computation*, eds A. Gopnik and L. Schultz (Oxford: Oxford University Press), 154–172.

LeBlanc, M. D., and Weber-Russell, S. (1996). Text integration and mathematical connections: a computer model of arithmetic word problem solving. *Cogn. Sci.* 20, 357–407. doi: 10.1207/s15516709cog2003_2

Lesage, E., Navarrete, G., and De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: the role of general cognitive resources. *Think. Reason.* 19, 27–53. doi: 10.1080/13546783.2012.713177

Lewis, C., and Keren, G. (1999). On the difficulties underlying Bayesian reasoning: comment on Gigerenzer and Hoffrage. *Psychol. Rev.* 106, 411–416. doi: 10.1037/0033-295X.106.2.411

Lipkus, I. M., and Peters, E. (2009). Understanding the role of numeracy in health: proposed theoretical framework and practical insights. *Health Educ. Behav.* 36, 1065–1081. doi: 10.1177/1090198109341533

Lipkus, I. M., Samsa, G., and Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Med. Decis. Making* 21, 37–44. doi: 10.1177/0272989X0102100105

Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. *Q. J. Exp. Psychol.* 48A, 188–207. doi: 10.1080/14640749508401384

Macchi, L., and Mosconi, G. (1998). Computational features vs. frequentist phrasing in the base-rate fallacy. *Swiss J. Psychol.* 57, 79–85.

Mandel, D. R. (2014a). The psychology of Bayesian reasoning. *Front. Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

Mandel, D. R. (2014b). Visual representation of rational belief revision: another look at the Sleeping Beauty problem. *Front. Psychol.* 5:1232. doi: 10.3389/fpsyg.2014.01232

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Front. Psychol.* 6:387. doi: 10.3389/fpsyg.2015.00387

Mayer, R. E. (1982). Memory for algebra story problems. *J. Exp. Psychol.* 74, 199–216. doi: 10.1037/0022-0663.74.2.199

Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instr. Sci.* 26, 49–63. doi: 10.1023/A:1003088013286

Mayer, R. E. (2003). "Mathematical problem solving," in *Mathematical Cognition: A Volume in Current Perspectives on Cognition, Learning, and Instruction*, ed J. M. Royer (Greenwich, CT: Information Age Publishing), 69–92.

Mayer, R. E., Lewis, A. B., and Hegarty, M. (1992). "Mathematical misunderstandings: qualitative reasoning about quantitative problems,"

in *The Nature and Origins of Mathematical Skills,* ed J. I. D. Campbell (Amsterdam: Elsevier Science Publishers), 137–154.

McNair, S., and Feeney, A. (2014). Does information about causal structure improve statistical reasoning? *Q. J. Exp. Psychol.* 67, 625–645. doi: 10.1080/17470218.2013.821709

McNair, S., and Feeney, A. (2015). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychon. Bull. Rev.* 22, 258–264. doi: 10.3758/s13423-014-0645-y

McNair, S. J. (2015). Beyond the status-quo: research on Bayesian reasoning must develop in both theory and method. *Front. Psychol.* 6:97. doi: 10.3389/fpsyg.2015.00097

Mellers, B. A., and McGraw, A. P. (1999). How to improve Bayesian reasoning: comment on Gigerenzer and Hoffrage (1995). *Psychol. Rev.* 106, 417–424. doi: 10.1037/0033-295X.106.2.417

Morsanyi, K., Primi, C., Chiesi, F., and Handley, S. (2009). The effects and side-effects of statistics education: psychology students' (mis-)conceptions of probability. *Contemp. Educ. Psychol.* 34, 210–220. doi: 10.1016/j.cedpsych.2009.05.001

Morsanyi, K. and Szücs, D. (2015). "Intuition in mathematical and probabilistic reasoning," in *The Oxford Handbook of Numerical Cognition*, eds R. Cohen Kadosh and A. Dowker (Oxford: Oxford University Press), 1–18. doi: 10.1093/oxfordhb/9780199642342.013.016

Murray, J. D., Klin, C. M., and Myers, J. L. (1993). Forward inferences in narrative text. *J. Mem. Lang.* 32, 464–473. doi: 10.1006/jmla.1993.1025

Navarrete, G., Correia, R., and Froimovitch, D. (2014). Communicating risk in prenatal screening: the consequences of Bayesian misapprehension. *Front. Psychol.* 5:1272. doi: 10.3389/fpsyg.2014.01272

Navarrete, G., and Santamaría, C. (2011). Ecological rationality and evolution: the mind really works that way? *Front. Psychol.* 2:251. doi: 10.3389/fpsyg.2011.00251

Nesher, P., and Teubal, E. (1975). Verbal cues as an interfering factor in verbal problem solving. *Educ. Stud. Math.* 6, 41–51. doi: 10.1007/BF00590023

Newell, A., and Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newell, B., and Hayes, B. (2007). Naturally nested, but why dual process? *Behav. Brain Sci.* 30, 276–277. doi: 10.1017/S0140525X07001847

Ni, Y., and Zhou, Y. D. (2005). Teaching and learning fraction and rational numbers: the origins and implications of whole number bias. *Educ. Psychol.* 40, 27–52. doi: 10.1207/s15326985ep4001_3

Oberauer, K. (2006). Reasoning with conditionals: a test of formal models of four theories. *Cogn. Psychol.* 53, 238–283. doi: 10.1016/j.cogpsych.2006.04.001

Oberauer, K., Hornig, R., Weidenfeld, A., and Wilhelm, O. (2005). Effects of directionality in deductive reasoning: II. Premise integration and conclusion evaluation. *Q. J. Exp. Psychol.* 58A, 1225–1247. doi: 10.1080/02724980443000566

Oberauer, K., and Wilhelm, O. (2000). Effects of directionality in deductive reasoning: I. The comprehension of single relational premises. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 1702–1712. doi: 10.1037/0278-7393.26.6.1702

Pacini, R., and Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *J. Pers. Soc. Psychol.* 76, 972–987. doi: 10.1037/0022-3514.76.6.972

Passolungh, M. C., and Pazzaglia, F. (2004). Individual differences in memory updating in relation to arithmetic problem solving. *Learn. Individ. Differ.* 14, 219–230. doi: 10.1016/j.lindif.2004.03.001

Paulos, J. A. (1988). *Innumeracy: Mathematical Illiteracy and Its Consequences*. New York, NY: Vintage Books.

Peters, E. (2012). Beyond comprehension : the role of numeracy in judgments and decisions. *Curr. Dir. Psychol. Sci.* 21, 31–35. doi: 10.1177/0963721411429960

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H., and Mertz, C. K. (2007). Less is more in presenting quality information to consumers. *Med. Care Res. Rev.* 64, 169–190. doi: 10.1177/1077558707070640020301

Sedlmeier, P., and Gigerenzer, G. (2001).Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Raven, J. C., Court, J. H., and Raven, J. (1977). *Manual for Advanced Progressive Matrices (Sets I and IT)*. London: H. K. Lewis and Co.

Reyna, V. F., and Brainerd, C. J. (2007). The importance of mathematics in health and human judgment: numeracy, risk communication, and medical decision making. *Learn. Individ. Differ.* 17, 147–159. doi: 10.1016/j.lindif.2007.03.010

Reyna, V. F., and Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learn. Individ. Differ.* 18, 89–107. doi: 10.1016/j.lindif.2007.03.011

Reyna, V. F., Lloyd, F. J., and Brainerd, C. J. (2003). "Memory, development, and rationality: an integrative theory of judgment and decision-making," in *Emerging Perspectives on Judgment and Decision Research,* eds S. Schneider and J. Shanteau (New York, NY: Cambridge University Press), 201–245.

Reyna, V. F., Nelson, W. L., Han, P. K., and Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychol. Bull.* 135, 943–973. doi: 10.1037/a0017327

Schoenfeld, A. (1985). *Mathematical Problem Solving*. New York, NY: Academic Press.

Schoenfeld, A. H. (1992). "Learning to think mathematically: problem solving, metacognition, and sense-making in mathematics," in *Handbook for Research on Mathematics Teaching and Learning,* ed D. Grouws (New York, NY: MacMillan), 334–370.

Schwartz, N., Strack, E., Hilton, D., and Naderer, G. (1991). Base rates, representativeness, and the logic of conversation: the contextual relevance of "irrelevant" information. *Soc. Cogn.* 9, 67–84. doi: 10.1521/soco.1991.9.1.67

Siegler, R. S., Fazio, L. K., Bailey, D. H., and Zhou, X. (2013). Fractions: the new frontier for theories of numerical development. *Trends Cogn. Sci.* 17, 13–19. doi: 10.1016/j.tics.2012.11.004

Siegler, R. S., Thompson, C. A., and Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cogn. Psychol.* 62, 273–296. doi: 10.1016/j.cogpsych.2011.03.001

Siegrist, M., and Keller, C. (2011). Natural frequencies and Bayesian reasoning: the impact of formal education and problem context. *J. Risk Res.* 14, 1039–1055. doi: 10.1080/13669877.2011.571786

Sirota, M., and Juanchich, M. (2011). Role of numeracy and cognitive reflection in Bayesian reasoning with natural frequencies. *Stud. Psychol.* 53, 151–161.

Sirota, M., Juanchich, M., and Hagmayer, Y. (2014a). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychon. Bull. Rev.* 21, 198–204. doi: 10.3758/s13423-013-0464-6

Sirota, M., Kostovičová, L., and Juanchich, M. (2014b). The effect of iconicity of visual displays on statistical reasoning: evidence in favor of the null hypothesis. *Psychon. Bull. Rev.* 21, 961–968. doi: 10.3758/s13423-013-0555-4

Sirota, M., Kostovičová, L., and Vallée-Tourangeau, F. (2015a). Now you Bayes, now you don't: effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychon. Bull. Rev.* doi: 10.3758/s13423-015-0810-y

Sirota, M., Kostovičová, L., and Vallée-Tourangeau, F. (2015b). How to train your Bayesian: a problem-representation transfer rather than a format-representation shift explains training effects. *Q. J. Exp. Psychol.* 68, 1–9. doi: 10.1080/17470218.2014.972420

Sirota, M., Vallée-Tourangeau, G., Vallée-Tourangeau, F., and Juanchich, M. (2015c). On Bayesian problem-solving: helping Bayesians solve simple Bayesian word problems. *Front. Psychol.* 6:1141. doi: 10.3389/fpsyg.2015.01141

Sloman, S. A., and Lagnado, D. (2015). Causality in thought. *Annu. Rev. Psychol.* 66, 3.1–3.25. doi: 10.1146/annurev-psych-010814-015135

Sloman, S. A., Over, D., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Process.* 91, 296–309. doi: 10.1016/S0749-5978(03)00021-9

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., and Voss, J. F. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *J. Verbal Learn. Verbal Behav.* 18, 275–290. doi: 10.1016/S0022-5371(79)90155-5

Stanovich, K. E. (2009). "Is it time for a tri-process theory. Distinguishing the reflective and the algorithmic mind," in *In Two Minds: Dual Processes and Beyond*, eds J. St. B. T. Evans and K. Frankish (Oxford: Oxford University Press), 55–88.

Stanovich, K. E., West, R. F., and Toplak, M. E. (2011). "Individual differences as essential components of heuristics and biases research," in *The Science of*

Reason, eds K. Manktelow, D. Over, and S. Elqayam (New York, NY: Psychology Press), 355–396.

Stupple, E. J. N., Ball, L. J., and Ellis, D. (2013). Matching bias in syllogistic reasoning: evidence for a dual-process account from response times and confidence ratings. *Think. Reason.* 19, 54–77. doi: 10.1080/13546783.2012.735622

Swanson, H. L., and Sachse-Lee, C. (2001). Mathematical problem solving and working memory in children with learning disabilities: both executive and phonological processes are important. *J. Exp. Child Psychol.* 79, 294–321. doi: 10.1006/jecp.2000.2587

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788

Thevenot, C., Barrouillet, P., and Fayol, M. (2004). Mental representation and procedures in arithmetic word problems: the effect of the position of the question. *Année Psychol.* 104, 683–699. doi: 10.3406/psy.2004.29685

Thevenot, C., and Oakhill, J. (2006). Representations and strategies for solving dynamic and static arithmetic word problems: the role of working memory capacities. *Eur. J. Cogn. Psychol.* 18, 756–775. doi: 10.1080/09541440500412270

Thompson, V. A. (2009). "Dual process theories: a metacognitive perspective," in *Two Minds: Dual Processes and Beyond,* eds J. Evans and K. Frankish (Oxford: Oxford University Press), 171–195.

Thompson, V. A. (2013). Why it matters: the implications of autonomous processes for dual process theories—commentary on Evans and Stanovich (2013). *Perspect. Psychol. Sci.* 8, 253–256. doi: 10.1177/1745691613483476

Tronsky, L. N., and Royer, J. M. (2003). "Relationships among basic computational automaticity, working memory, and complex mathematical problem solving," in *Mathematical Cognition,* ed J. M. Royer (Greenwich, CT: Information Age Publishing), 117–146.

Tubau, E. (2008). Enhancing probabilistic reasoning: the role of causal graphs, statistical format and numerical skills. *Learn. Individ. Dif.* 18, 187–196. doi: 10.1016/j.lindif.2007.08.006

Tubau, E., Aguilar-Lleyda, D., and Johnson, E. D. (2015). Reasoning and choice in the Monty Hall Dilemma (MHD): implications for improving Bayesian reasoning. *Front. Psychol.* 6:353. doi: 10.3389/fpsyg.2015.00353

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90, 293–315. doi: 10.1037/0033-295X.90.4.293

Vallée-Tourangeau, G., Abadie, M., and Vallee-Tourangeau, F. (2015a). Interactivity fosters Bayesian reasoning without instruction. *J. Exp. Psychol. Gen.* 144, 581–603. doi: 10.1037/a0039161

Vallée-Tourangeau, G., Sirota, M., Juanchich, M., and Vallée-Tourangeau, F. (2015b). Beyond getting the numbers right: what does it mean to be a 'successful' Bayesian reasoner? *Front. Psychol.* 6:712. doi: 10.3389/fpsyg.2015.00712

van den Broek, P. (1990). Causal inferences and the comprehension of narrative texts. *Psychol. Learn. Motiv.* 25, 175–196. doi: 10.1016/S0079-7421(08)60255-8

van Dijk, T. A., and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York, NY: Academic Press.

Verschaffel, L., De Corte, E., and Pauwels, A. (1992). Solving compare problems: an eye movement test of Lewis and Mayer's consistency hypothesis. *J. Educ. Psychol.* 84, 85–94. doi: 10.1037/0022-0663.84.1.85

Villejoubert, G., and Mandel, D. R. (2002). The inverse fallacy: an account of deviations from Bayes's theorem and the additivity principle. *Mem. Cognit.* 30, 171–178. doi: 10.3758/BF03195278

Vinner, S. (1997). The pseudo-conceptual and the pseudo-analytical thought processes in mathematics learning. *Educ. Stud. Math.* 34, 97–129. doi: 10.1023/A:1002998529016

Vranas, P. B. M. (2000). Gigerenzer's normative critique of Kahneman and Tversky. *Cognition* 76, 179–193. doi: 10.1016/S0010-0277(99)00084-0

Waldmann, M. R., Hagmayer, Y., and Blaisdell, A. P. (2006). Beyond the information given. *Curr. Dir. Psychol. Sci.* 15, 307–311. doi: 10.1111/j.1467-8721.2006.00458.x

Wolfe, C. R. (1995). Information seeking on Bayesian conditional probability problems: a fuzzy-trace theory account. *J. Behav. Decis. Mak.* 8, 85–108. doi: 10.1002/bdm.3960080203

Wolpert, D. M., and Ghahramani, Z. (2005). "Bayes rule in perception, action and cognition," in *Oxford Companion to Consciousness,* ed R. L. Gregory (New York, NY: Oxford University Press), 1–4.

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: frequency or nested sets? *Exp. Psychol.* 50, 97–106. doi: 10.1026/1618-3169.50.2.97

Zhu, L., and Gigerenzer, G. (2006). Children can solve Bayesian problems: the role of representation in mental computation. *Cognition* 98, 287–308. doi: 10.1016/j.cognition.2004.12.003

Zukier, H., and Pepitone, A. (1984). Social roles and strategies in prediction: some determinants of the use of base-rate information. *J. Pers. Soc. Psychol.* 47, 349–360. doi: 10.1037/0022-3514.47.2.349

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

OPINION
published: 10 August 2015
doi: 10.3389/fpsyg.2015.01141

On Bayesian problem-solving: helping Bayesians solve simple Bayesian word problems

Miroslav Sirota[1]*, *Gaëlle Vallée-Tourangeau*[1], *Frédéric Vallée-Tourangeau*[1] and *Marie Juanchich*[2]

[1] Department of Psychology, Kingston University, London, UK, [2] Department of Management, Kingston University, London, UK

Keywords: Bayesian problem-solving, Bayesian research paradox, natural frequencies, Bayesian reasoning, mathematical problem solving

Resolving the "Bayesian Paradox"—Bayesians Who Failed to Solve Bayesian Problems

A well-supported conclusion a reader would draw from the vast amount of research on Bayesian inference could be distilled into one sentence: "People are profoundly Bayesians, but they fail to solve Bayesian word problems." Indeed, two strands of research tell different stories about our ability to make Bayesian inferences—our ability to infer posterior probability from prior probability and new evidence according to Bayes's theorem. People see, move, coordinate, remember, learn, reason and argue consistently with complex probabilistic Bayesian computations, but they fail to solve, computationally much simpler, Bayesian word problems.

On the one hand, a first strand of research shows that people are profoundly Bayesians. Strong evidence indicates that the brain represents probability distributions and certain neural circuits perform Bayesian computations (Pouget et al., 2013). Bayesian computation models account for a wide range of observations on sensory perception, motoric behavior and sensorimotor coordination (see Chater et al., 2010; Pouget et al., 2013). Bayesian computations approximate observed patterns in inductive reasoning, memory, language production, and language comprehension (Chater et al., 2010). Even 12-month-old preverbal infants present behavior consistent with the behavior of a Bayesian ideal observer: infants integrate multiple sources of information to form rational expectations about situations they have never encountered before (Téglás et al., 2011). In everyday life, people form cognitive judgments predicting the occurrence of everyday events consistent with a Bayesian ideal observer (Griffiths and Tenenbaum, 2006).

On the other hand, however, a second strand of research shows that people fail to make the simplest possible Bayesian inference once they are presented with Bayesian word problems. Indeed, people tend to largely ignore or neglect base-rate information in probability judgment tasks such as social judgment or textbook problem tasks (Kahneman and Tversky, 1973; Bar-Hillel, 1980) or they tend to fail to be Bayesians in a completely opposite way—by overweighting base-rate information (Teigen and Keren, 2007). In fact, people require costly and intense training with most statistical formats to achieve good performance with probabilistic inferences that deteriorates with time very quickly (Sedlmeier and Gigerenzer, 2001).

So people are Bayesians who fail to solve simple Bayesian word problems. As with most paradoxes, a solution to this "Bayesian paradox" lies in taking closer look at conceptualizations: at what constitutes a Bayesian inference in these two strands of research. Such analysis uncovers important design differences, Bayesian classification criteria and statistical approaches (Vallée-Tourangeau et al., 2015). However, the crucial difference that we

OPEN ACCESS

Edited by:
David R. Mandel, Defence Research and Development Canada, Canada

Reviewed by:
W. Trey Hill, Fort Hays State University, USA
Ulrich Hoffrage, Université de Lausanne, Switzerland

***Correspondence:**
Miroslav Sirota, miroslav.sirota@kingston.ac.uk

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *31 March 2015*
Accepted: *22 July 2015*
Published: *10 August 2015*

Citation:
Sirota M, Vallée-Tourangeau G, Vallée-Tourangeau F and Juanchich M (2015) On Bayesian problem-solving: helping Bayesians solve simple Bayesian word problems. Front. Psychol. 6:1141. doi: 10.3389/fpsyg.2015.01141

highlight here lies in the cognitive processes involved in performing the task. What is described as a "Bayesian inference" in the two strands conflates very different processes. *Implicit processes*—implicit calculations with probabilities mostly acquired from experience—are involved in the Bayesian computations approximating the performance of various cognitive functions and in the estimation of experienced real-life outcomes. *Explicit processes*—explicit calculations with probabilities typically extracted from a textual description—are involved in solving Bayesian textbook problems or social judgment problems. The different information source, experience or description, for example, has been shown to lead to dramatically different choices and decisions (e.g., Hertwig et al., 2004). With this distinction, of course, we do not intend to imply that all the cognitive processes involved in estimating probabilities are necessarily implicit and engage only with the probabilities from experience or vice versa. Rather we wish to point out that the different experimental paradigms outlined here require typically different cognitive processes operating over different types of information.

This postulated distinction between cognitive processes involved in these different types of Bayesian inference tasks can be mapped onto a distinction between biologically primary (pan-cultural, evolutionary purposeful) cognitive abilities and biologically secondary (culturally specific) cognitive abilities (Geary, 1995). It could also be linked to the debate on how people form probability judgments, either through automatic frequency encoding of sequentially presented information (e.g., Hasher and Zacks, 1984) or through heuristic inferences from aggregated information (e.g., representativeness heuristic, Kahneman and Tversky, 1974).

Which type of evidence should we call upon to help us decide whether people are Bayesians or not? Both implicit and explicit processes are relevant for assessing this ability. Having Bayesian eyes, hands and minds is arguably important for survival. Yet, our environment has changed dramatically in the Twentieth century—it became crowded with explicit aggregated statistical information. Learning from described aggregated information condenses the learning process compared with learning from experience. Imagine, for example, an experienced UK physician relocating to Nigeria. Her experience would provide her with an adequate knowledge of the disease base rates, sensitivity and specificity of medical tests within the UK population; however her experience may not be applicable or may even be deleterious in Nigeria given that those pieces of information may differ. The doctor would greatly benefit from reading explicit aggregated statistical information on base rates of diseases, sensitivity and specificity of medical tests in the local population to avoid making errors and the long learning process based on personal experience. Most importantly, she should be able to integrate this information into her diagnostic judgments when facing a given set of symptoms in a patient in Nigeria. More generally, in their probability-laden environment, all people (not just physicians) may come across a lot of problems similar to Bayesian textbook problems, of which cancer or prenatal screening are just examples (e.g., Navarrete et al., 2014). It is clear, therefore, that we should focus on improving the explicit processes that underpin Bayesian reasoning as a problem-solving ability.

Bayesian Problem-solving

Although the processes involved in solving Bayesian textbook problems resemble the processes involved in solving other mathematical problems, research on Bayesian reasoning has evolved in parallel to the research on problem solving. Reframing processes involved in Bayesian textbook reasoning in terms of the processes examined in the problem-solving literature can benefit Bayesian reasoning research efforts. The problem-solving literature not only extends the sound methodological toolkit to explore underpinning mental processes (e.g., thinking aloud protocols), but it also offers alternative concepts enacting novel insights, different explanations and more elaborate models generating deeper understanding of Bayesian problem-solving. We outline three examples of such theoretical benefits in the context of facilitating Bayesian problem-solving.

First, applying problem-solving concepts to Bayesian reasoning offers a novel and productive perspective. For example, we could think of Bayesian textbook problems in a problem-solving framework as a combination of insight and analytical problems. Typically, the problem-solving literature distinguishes two classes of problems: analytical and insight problems (Gilhooly and Murphy, 2005). With analytical problems, people can work out an incremental solution and rarely experience an Aha! moment in the process. Consider, for instance, this multi-digit addition problem: "Sum up the following numbers: 13, 27, 12, 32, 25, 11"; participants announcing an answer rarely do so with Eureka glee (although they might experience relief). With insight problems, people have to overcome an initial impasse to reach a completely new way of thinking about the problem; they need to transform the initial problem representation into a new representation which will lead them to the goal state. Consider, for instance, the following problem: "Place 17 animals in 4 enclosures in such a manner that there will be an odd number of animals in each enclosure" (adapted from Metcalfe and Wiebe, 1987). You probably try 17/4 and it did not work: The problem masquerades as an arithmetic puzzle. However, in contrast to an analytic problem, the initial problem presentation cannot be transformed step-by-step to a solution (in this case the solution involves overlapping sets). This distinction suggests that decomposing the question of "What facilitates Bayesian reasoning?" into "What facilitates the insight?" and "What facilitates the computation?" will pave the way for better understanding what factors facilitate the problem-structure understanding and what factors facilitate the computational operations in Bayesian problem-solving (see also Johnson and Tubau, 2015).

Second, rephrasing Bayesian reasoning as a form of problem-solving offers different explanations of the processes implicated, for example, those involved in representational training (e.g., Sedlmeier and Gigerenzer, 2001; Mandel, 2015; Sirota et al., 2015a). In representational training, participants learn to transform the statistical format representation of a problem—they learn to translate single-event probabilities into natural

frequencies. For example, the statements "a 1% probability that a woman has breast cancer" and "if a woman has cancer then there is an 80% probability that she will get a positive mammogram" are translated as "10 out of every 1000 women have breast cancer" and "8 out of the 10 who have breast cancer will get a positive mammogram." The problem-solving approach posits that the underlying mechanism of such representational training consists of the acquisition of an appropriate problem representation—a nested-sets representation of the Bayesian problem, regardless of frequencies or probabilistic information contained in such problem—during the learning phase, which is then transferred to similar problems in the testing phase (for evidence see Sirota et al., 2015a). This goes beyond the default explanation that participants translated single-event probabilities into natural frequencies (Sedlmeier and Gigerenzer, 2001) and it accounts for the training success in terms of the specific mental processes involved in problem representation learning and its transfer (for the importance of a good representation in different problems of a belief revision not depending on natural frequencies, see Mandel, 2014).

Third, recruiting problem-solving models offers a better understanding of well-known effects in Bayesian reasoning than we currently have, for example, the format effect. Statistical formats such as natural frequencies represent probably the most cost-effective (and the most discussed) tool to facilitate Bayesian problem-solving, given that visual aids offer mixed evidence of their effectiveness (e.g., Cosmides and Tooby, 1996; Sirota et al., 2014b). Natural frequencies enhance Bayesian problem-solving when compared with formats involving normalization such as probability formats (e.g., Gigerenzer and Hoffrage, 1995; Cosmides and Tooby, 1996; Barbey and Sloman, 2007). Natural frequencies, introduced by Kleiter (1994), integrate the base-rate information in their structure making the base-rate information per se redundant. For example, the statement "8 women out of the 10 who have breast cancer will get a positive mammogram" includes the base-rate information of *the* 10 (out of 1000) women with cancer from our previous example.

According to the general framework of mathematical verbal problem solving (Kintsch and Greeno, 1985; Kintsch, 1988), which integrates formal mathematical and linguistic knowledge, two processes should be differentiated here: the processes involved in representing the problem and those involved in producing a solution (for specific approaches to probability representation, see Johnson-Laird et al., 1999; Mandel, 2008). In the problem representation phase, a mental representation is constructed from the text that triggers available knowledge schemas stored in long-term memory. Familiar cues in the text activate a correct mental representation of the problem more easily than unfamiliar or misleading ones; this enables an easier integration with existing knowledge. In the problem solution phase, rules or strategies corresponding to the problem representation are implemented. We suggest that the facilitative effect of natural frequencies in Bayesian inference problems is due to a similar process. A wording of the task with frequencies (e.g., explicit set reference language such as "10 *out of* the *remaining* 90")—not the numerical format by itself—may trigger a representation of the problem as nested sets, while a wording of the task with probabilities which conceal the nested set structure due to normalizing, does not. Such an explanation casts natural frequencies as a familiar format rather than a privileged one. Some authors view natural frequencies as a privileged format because they are processed by a specialized frequency-coding mechanism shaped by evolutionary forces (Gigerenzer and Hoffrage, 1995). If true (and some specific conditions are fulfilled, Barrett et al., 2006) then processing of a privileged format should not be cognitively demanding at all or at least less cognitively demanding than processing of a computationally equivalent and equally familiar format (e.g., Cosmides and Tooby, 1996). It means, for instance, that measures of cognitive capacity should not be predictive of performance in Bayesian reasoning. However, several recent studies have provided evidence rebutting the claim of easier processing of natural frequencies (Sirota and Juanchich, 2011; Lesage et al., 2013; Sirota et al., 2014a).

Conclusion

Our environment is laden with statistical information and demands from people that they successfully solve problems that are exactly the same as, or similar to, classical Bayesian textbook problems. Although some brain function appears to implement Bayesian computations, people's abilities to solve Bayesian word problems could still be substantially improved. We should therefore strive to understand and improve people's performance with this kind of problems. We suggest thinking about the involved processes as processes akin to those engaged during problem-solving (see also Johnson and Tubau, 2015; Sirota et al., 2015b). Such a re-classification would not only resolve contradictions in research on Bayesian inference, it would also facilitate the application of conceptual and methodological tools from problem-solving research. It would allow us to ask what enacts the insight about the problem structure, what facilitates the relevant computations and how exactly people implement these processes. It would allow us to conceptually re-frame observed effects such as representational training effects. It would also allow us to shed more light on the underlying processes by utilizing elaborate process-oriented models developed in this area.

Acknowledgments

We thank David Mandel, Ulrich Hoffrage and the anonymous reviewer for helpful suggestions on an earlier version of this manuscript.

References

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

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–254. doi: 10.1017/S0140525X07001653

Barrett, H. C., Frederick, D. A., Haselton, M. G., and Kurzban, R. (2006). Can manipulations of cognitive load be used to test evolutionary hypotheses? *J. Pers. Soc. Psychol.* 91, 513. doi: 10.1037/0022-3514.91.3.513

Chater, N., Oaksford, M., Hahn, U., and Heit, E. (2010). Bayesian models of cognition. *Wiley Interdiscipl. Rev. Cogn. Sci.* 1, 811–823. doi: 10.1002/wcs.79

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

Geary, D. C. (1995). Reflections of evolution and culture in children's cognition: implications for mathematical development and instruction. *Am. Psychol.* 50, 24. doi: 10.1037/0003-066X.50.1.24

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction - Frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gilhooly, K. J., and Murphy, P. (2005). Differentiating insight from non-insight problems. *Think. Reason.* 11, 279–302. doi: 10.1080/13546780442000187

Griffiths, T. L., and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychol. Sci.* 17, 767–773. doi: 10.1111/j.1467-9280.2006.01780.x

Hasher, L., and Zacks, R. T. (1984). Automatic processing of fundamental information: the case of frequency of occurrence. *Am. Psychol.* 39, 1372. doi: 10.1037/0003-066x.39.12.1372

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychol. Sci.* 15, 534–539. doi: 10.1111/j.0956-7976.2004.00715.x

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., and Caverni, J. P. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychol. Rev.* 106, 62–88. doi: 10.1037/0033-295X.106.1.62

Johnson, E. D., and Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Front. Psychol.* 6:938. doi: 10.3389/fpsyg.2015.00938

Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychol. Rev.* 80, 237. doi: 10.1037/h0034747

Kahneman, D., and Tversky, A. (1974). "Subjective probability: a judgment of representativeness," in *The Concept of Probability in Psychological Experiments,* ed C.-A. S. Staël Von Holstein (Dordrecht: Springer), 25–48.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. Rev.* 95, 163. doi: 10.1037/0033-295X.95.2.163

Kintsch, W., and Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychol. Rev.* 92, 109. doi: 10.1037/0033-295X.92.1.109

Kleiter, G. D. (1994). "Natural sampling: rationality without base rates," in *Contributions to Mathematical Psychology, Psychometrics, and Methodology,* eds G. Fischer and D. Laming (New York, NY: Springer), 375–388. doi: 10.1007/978-1-4612-4308-3_27

Lesage, E., Navarrete, G., and De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: the role of general cognitive resources. *Think. Reason.* 19, 27–53. doi: 10.1080/13546783.2012.713177

Mandel, D. R. (2008). Violations of coherence in subjective probability: a representational and assessment processes account. *Cognition* 106, 130–156. doi: 10.1016/j.cognition.2007.01.001

Mandel, D. R. (2014). Visual representation of rational belief revision: another look at the Sleeping Beauty problem. *Front. Psychol.* 5:1232. doi: 10.3389/fpsyg.2014.01232

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Front. Psychol.* 6:387. doi: 10.3389/fpsyg.2015.00387

Metcalfe, J., and Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Mem. Cogn.* 15, 238–246. doi: 10.3758/BF03197722

Navarrete, G., Correia, R., and Froimovitch, D. (2014). Communicating risk in prenatal screening: the consequences of Bayesian misapprehension. *Front. Psychol.* 5:1272. doi: 10.3389/fpsyg.2014.01272

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* 16, 1170–1178. doi: 10.1038/nn.3495

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Sirota, M., and Juanchich, M. (2011). Role of numeracy and cognitive reflection in Bayesian reasoning with natural frequencies. *Stud. Psychol.* 53, 151–161.

Sirota, M., Juanchich, M., and Hagmayer, Y. (2014a). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychon. Bull. Rev.* 21, 198–204. doi: 10.3758/s13423-013-0464-6

Sirota, M., Kostovičová, L., and Juanchich, M. (2014b). The effect of iconicity of visual displays on statistical reasoning: evidence in favor of the null hypothesis. *Psychon. Bull. Rev.* 21, 961–968. doi: 10.3758/s13423-013-0555-4

Sirota, M., Kostovičová, L., and Vallee-Tourangeau, F. (2015a). How to train your Bayesian: A problem-representation transfer rather than a format-representation shift explains training effects. *Q. J. Exp. Psychol.* 68, 1–9. doi: 10.1080/17470218.2014.972420

Sirota, M., Kostovičová, L., and Vallée-Tourangeau, F. (2015b). Now you Bayes, now you don't: effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychon. Bull. Rev.* doi: 10.3758/s13423-015-0810-y. [Epub ahead of print].

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., and Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science* 332, 1054–1059. doi: 10.1126/science.1196404

Teigen, K. H., and Keren, G. (2007). Waiting for the bus: when base-rates refuse to be neglected. *Cognition* 103, 337–357. doi: 10.1016/j.cognition.2006.03.007

Vallée-Tourangeau, G., Sirota, M., Juanchich, M., and Vallée-Tourangeau, F. (2015). Beyond getting the numbers right: what does it mean to be a "successful" Bayesian reasoner? *Front. Psychol.* 6:712. doi: 10.3389/fpsyg.2015.00712

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in PSYCHOLOGY

OPINION ARTICLE
published: 04 February 2015
doi: 10.3389/fpsyg.2015.00070

Controlled information integration and bayesian inference

Peter Juslin*

Department of Psychology, Uppsala University, Uppsala, Sweden
**Correspondence: peter.juslin@psyk.uu.se*

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
Johan Kwisthout, Radboud University Nijmegen, Netherlands
Karin Binder, University of Regensburg, Germany

Keywords: linear additive integration, probability reasoning, base-rate neglect, working memory capacity, Bayesian inference

One of the oldest hypotheses in cognitive psychology is that controlled information integration[1] is a serial, capacity-constrained process that is delimited by our working memory resources, and this seems to be the most uncontroversial aspect also of present-day dual-systems theories (Evans, 2008). The process is typically conceived of as a sequential adjustment of an estimate of a criterion (e.g., a probability), in view of successive consideration of inputs to the judgment (i.e., cues or evidence). The "cognitive default" seems to be to consider each attended cue in isolation, taking its impact on the criterion into account by adjusting a previous estimate into a new estimate, until a stopping rule applies (e.g., Juslin et al., 2008).

Considering each input in isolation, without modifying the adjustments contingently on other inputs to the judgment, invites *additive integration*. The limits on working memory moreover contribute to an illusion of linearity. If people, when pondering the relationship between variables X and Y, are constrained by working memory to consider only two X–Y pairs, the function induced can take no other form than a line. As illustrated by many scientific models, with computational aids people can capture also non-additive and non-linear relations. But without support, this is rather taxing on working memory and additive integration, typically as a weighted average, seems to be the default process (Juslin et al., 2009), and, even more so, considering that additive integration is famously "robust" (Dawes, 1979), allowing little marginal benefit from also considering the putative configural effects of cues. These cognitive constraints therefore define a point toward which our judgments naturally gravitate.

This simplistic and probably not overly controversial model of controlled integration immediately has important consequences for our abilities to make judgments, some of which are well-known, some of which may still need to be further digested. At a general level, the most fundamental constraint on people's ability to comprehend and control their environment is this tendency to view it in terms of an "additive caricature," as if they "looked at the world through a straw," appreciating each factor in isolation, but with limited ability to capture the interactions and dynamics of the entire system. In more prosaic terms, a wealth of evidence suggests that multiple-cue judgments are typically well described by simple linear additive models (Brehmer, 1994; Karelaia and Hogarth, 2008), even if the task departs from linearity and additivity.

There are important exceptions where people transcend this imprisonment in a linear additive mental universe also without external computational aids, in particular, an ability to use a prior input to "contextualize" the meaning of an immediately following input. For example, for a lottery, like a 0.10 chance of winning \$100 and \$0 otherwise, people have little difficulty with contextualizing the outcome in view of the preceding probability; that is, to discount the "appeal" of the positive outcome of receiving \$100 by the fact that the probability of ever seeing it is low. Likewise, people often have little difficulty with understanding normalized probability ratios and appreciate that, say, "30 chances in 100" and "300 chances in 1000" describe comparable states of uncertainty, something that again requires that one input is contextualized by another[2]. These exceptions are important, but seem to be connected to specific judgment domains.

CONTROLLED INTEGRATION AND PROBABILITY THEORY

This contrasts with the requirements for multiplication implied by many rules of probability theory. We have therefore argued that additive combination may be an important—and often neglected—constraint on people's ability to reason with probability. Nilsson et al. (2009) proposed that even a classic bias, like the conjunction fallacy (Kahneman and Frederick, 2002), may not primarily be explained by specific heuristics *per se*, like "representativeness," as typically claimed (although people sometimes use representativeness to make these judgments), but by a tendency to combine constituent probabilities by additive combination (see also Nilsson et al., 2013, 2014; Jenny et al., 2014). For example, people may appreciate that a description of "Linda" is likely if she is a feminist and unlikely if she is a bank teller (which might be mediated by "representativeness"), but knowing no feminist bank tellers they combine these assessments as best they can, which typically comes out as a weighted average (Nilsson et al., 2009). The rate of conjunction errors

[1]Controlled processes refer to cognitive processes that are slow, conscious, intentional, and constrained by attention, in contrast to automatic processes that are rapid, not constrained by attention, and can be triggered also directly by stimulus properties (Schneider and Shiffrin, 1977; see also Evans, 2008). The claims about cognitive constraints discussed in this article refer to controlled processes and automatic processes may often better approximate Bayesian information integration (see, e.g., Tenenbaum et al., 2011).

[2]This ability is not perfect as illustrated by the phenomenon of denominator neglect (Reyna and Brainerd, 2008).

indeed seems equally high regardless of whether the representativeness heuristic is applicable or not (Gavanski and Roskos-Ewoldsen, 1991; Nilsson, 2008).

Juslin et al. (2011) similarly argued that base-rate neglect may be explained not by use of specific heuristics *per se*, but by additive combination of base-rates, hit-rates, and false alarm rates, where the weighting of the components is context-dependent (and more often neglect false-alarm rates than base-rates)[3]. Importantly, the reliance on additive integration is by no means arbitrary: to the extent that people base their judgments on noisy input (e.g., small samples), linear additive integration often yields as accurate judgments as reliance on probability theory, possibly explaining why the mind has evolved with little appreciation for the integration implied by probability theory (Juslin et al., 2009).

A strong example of problems with probability integration comes from studies of experienced bettors that have played on soccer games at least a couple of times each month for a period of 10 years or more (Nilsson and Andersson, 2010; Andersson and Nilsson, in press). They were extremely accurate in their translation of odds into probabilities, including that they aptly captured the profit margin introduced in the odds by the gambling companies. Yet, when they assessed the odds of an unlikely event *A* (i.e., an outcome of a soccer game), the odds for the conjunction of *A* and a likely event *B*, and the odds of the conjunction of *A*, *B*, and a third likely event *C*, their probability assessments and their willingness to pay for the bet, increased as likely events were added to the conjunction (the conjunction fallacy). This is predicted by a weighted average of the components, but violates probability theory. Exquisite assessment, but blatantly "irrational" integration, also in experienced and very motivated probability reasoners.

[3]A linear additive model captures many properties of the data, such that people do appreciate the qualitative effect of the base-rate, flexibly change their weighting as a function of contextual cues, and that the judgments are typically less extreme as compared to Bayes' theorem, but until we have a theory of how contextual cues affect the weight of the base-rate, we have limited ability to predict a priori how the base-rate will be used in a specific situation.

BAYESIAN INFERENCE

Bayes' theorem in its odds format is,

$$p(H|E)/p(-H|E) = p(H)/p(-H) \cdot p(E|H)/p(E|-H) \quad (1)$$

where the left-hand side is the posterior odds for hypothesis *H* given evidence *E*, the first right-hand component is the prior odds for hypothesis *H*, and the second right-hand side is the likelihood ratio for the evidence *E*, given that *H* is true or false (i.e., $-H$). Equation (1) can be used to adjust your subjective probability that hypothesis *H* is true, in the light of evidence *E*.

Although apparently simple, the adjustment of the probability required in view of the evidence depends not only on the evidence attended at the moment, but on the prior probability (e.g., when the likelihood ratio is 2, you should adjust the prior probability of *H* upwards by 0.17 if the prior ratio is 1, but upwards by 0.04 if the prior ratio is 10)[4]. People do appreciate that the posterior probability is a positive function both of the prior and the evidence, but the impact of the prior is typically less than expected from Bayes' theorem (Koehler, 1996). If people, as argued above, are spontaneously inclined to adjust the probability of *H* (criterion) in the light of the new evidence *E* (the currently attended cue) independently of the previous input (captured in the prior probability), they will be affected by both priors and evidence, but not as much as with Equation (1), because they combine them additively [5]. This account explains why people find this a difficult task, but also suggests simplifying conditions and a "cure" for base-rate neglect.

A first example of a simplifying condition is natural frequencies (Gigerenzer and Hoffrage, 1995). If the base-rate problem immediately conveys the number of people with, say, a positive mammography test and the number of such people with breast cancer, people can "contextualize" the second number in terms of the first and directly appreciate that among positive tests, the proportion of breast cancer is low. In belief revision tasks, where the belief is repeatedly updated in the face of evidence, it has long been known that people successively average the "old" and "new" data (e.g., Shanteau, 1972; Lopes, 1985; Hogarth and Einhorn, 1992; McKenzie, 1994). An exception is when prior and evidence are presented in contextual and temporal contiguity, where people have some ability to "contextualize" their, presumably also here linear, weighting of the evidence in view of the prior, better emulating Bayesian integration (Shanteau, 1975).

The "cure" to base-rate neglect suggested by this view is, of course, to replace multiplicative integration with additive integration. An immediate implication is that people should have very little problem with certain kinds of "Bayesian updating;" for example, with updating their prior belief about the mean in a population after observing a new sample from the population. "Bayesian updating" here amounts to a (sample-size) weighted average between the "prior mean" and the "sample mean," a task that people should be able to learn quite easily.

An example directly related to Bayes' theorem is provided in Juslin et al. (2011). In Experiment 1, each participant responded to 30 medical diagnosis tasks, in one of three formats: (i) *standard probability*, The base-rate, hit-rate, and false alarm rate were stated as probabilities[6]; (ii) *odds*, The same problem expressed in prior odds and likelihood ratios (Equation 1); (iii) *Log odds*, The same problems expressed as log odds, implying that one simply adds the log prior odds to the log likelihood odds to arrive at the log posterior odds. These are three ways to represent the same problems, but the first two formats require multiplication, the last one additive

[4]With prior odds 1 and likelihood ratio 2, the posterior odds is 2 (Equation 1); an adjustment from a prior probability of 0.5 to a posterior probability of 0.67. With prior odds 10, the corresponding adjustment will be from 0.91 to 0.95.

[5]More specifically, when the base-rate is extreme, as in the "mammography problem" (e.g., Gigerenzer and Hoffrage, 1995) people will "underuse" the base-rate, but in problems with ambiguous base-rate, like in the urn problems studied by Edwards (1982), they will "overuse" the base-rate and thus appear "conservative."

[6]Here is an example of a medical diagnosis task: The probability that a person randomly selected from the population of all Swedes has the disease is 2%. The probability of receiving a positive test result given that one has the disease is 96%. The probability of receiving a positive test result if one does not have the disease is 8%. What is the probability that a randomly selected person with a positive test result has the disease? Correct answer: 20%.

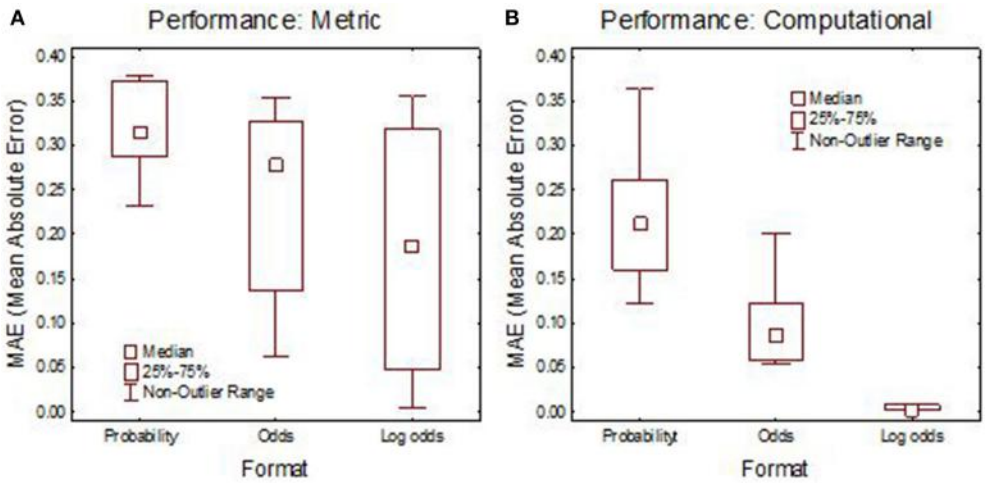

FIGURE 1 | Median performance in Experiment 1 in terms of Mean Absolute Error (MAE) between the judgment and Bayes' theorem. (A) Metric instruction; **(B)** computational instruction. Adapted from Juslin et al. (2011) with permission.

integration. Fifteen participants received *Metric instruction*, explaining and exemplifying the range and sign of the metric used, but with no guidance on how the integration should be made. The other 15, in addition, received *Computational instructions* on how to solve the problems, explaining how the components should be integrated according to Bayes' theorem with numerical examples.

The performance is summarized in **Figure 1**. Already with a Metric instruction, the log-odds format produced judgments closer to Bayes' theorem than the standard probability format. With computational instruction, the standard probability format produced poor performance and participants were still better described by an additive than a multiplicative (Bayesian) model. With log odds and computational instruction, performance was in perfect agreement with Bayes' theorem. People can thus flawlessly perform Bayesian calculation when the integration is additive, but when the format requires multiplication they are inept also after explicit instruction, still approximating Bayes' theorem as best they can by a linear additive combination.

CONCLUSIONS

A caveat is that although these results demonstrate limits on *computational ability*, admittedly they do not address the important issue of *computational insight*: the understanding of what needs to be computed in the first place. Research has emphasized conditions that foster computational insight by highlighting subset relations that are important in Bayesian reasoning problems (e.g., Barbey and Sloman, 2007), perhaps at the neglect of the "old-school" information processing constraints on people's computational abilities discussed here. The "cure" suggested here is drastic in the sense that it requires people to think of uncertainty in an unfamiliar log odds format, and the extent to which they can learn to do this is an open question. The dilemma might well be that the probability format is more easily translated into action, because probabilities can be used directly to fraction-wise "contextualize" (discount) decision outcomes, but for reasoning about uncertainty people are better off with formats that allow additive integration.

REFERENCES

Andersson, P., and Nilsson, H. (in press). Do bettors correctly perceive odds? Three studies of how bettors interpret betting odds as probabilistic information. *J. Behav. Decis. Making.*

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–254. doi: 10.1017/S0140525X07001653

Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychol.* 87, 137–154. doi: 10.1016/0001-6918(94)90048-5

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *Am. Psychol.* 34, 571–582. doi: 10.1037/0003-066X.34.7.571

Edwards, W. (1982). "Conservatism in human information processing," in *Judgment Under Uncertainty: Heuristics and Biases,* eds D. Kahneman, P. Slovic, and A. Tversky (Cambridge: Cambridge University Press), 359–369. doi: 10.1017/CBO9780511809477.026

Evans, J. B. T. (2008). Dual processing accounts of reasoning judgment and social cognition. *Annu. Rev. Psychol.* 59, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Gavanski, I., and Roskos-Ewoldsen, D. R. (1991). Representativeness and conjoint probability. *J. Pers. Soc. Psychol.* 61, 181–194. doi: 10.1037/0022-3514.61.2.181

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Hogarth, R. M., and Einhorn, H. J. (1992). Order effects in belief updating: the belief-adjustment model. *Cogn. Psychol.* 24, 1–55. doi: 10.1016/0010-0285(92)90002-J

Jenny, M. A., Rieskamp, J., and Nilsson, H. (2014). Inferring conjunctive probabilities from noisy samples: evidence for the configural weighted average model. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 203–217. doi: 10.1037/a0034261

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

Juslin, P., Nilsson, H., and Winman, A. (2009). Probability theory: not the very guide of life. *Psychol. Rev.* 116, 856–874. doi: 10.1037/a0016979

Juslin, P., Nilsson, H., Winman, A., and Lindskog, M. (2011). Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats. *Cognition* 120, 248–267. doi: 10.1016/j.cognition.2011.05.004

Kahneman, D., and Frederick, S. (2002). "Representativeness revisited: attribute substitution in intuitive judgment," in *Heuristics and Biases: The Psychology of Intuitive Judgment,* eds T. Gilovich, D. W. Griffin, and D. Kahneman (New

York, NY: Cambridge University Press), 49–81. doi: 10.1017/CBO9780511808098.004

Karelaia, N., and Hogarth, R. M. (2008). Determinants of linear judgment: a meta-analysis of lens studies. *Psychol. Bull.* 134, 404–426. doi: 10.1037/0033-2909.134.3.404

Koehler, J. J. (1996). The base-rate fallacy reconsidered: descriptive, normative and methodological challenges. *Behav. Brain Sci.* 19, 1–17. doi: 10.1017/S0140525X00041157

Lopes, L. L. (1985). Averaging rules and adjustment processes in Bayesian inference. *Bull. Psychon. Soc.* 23, 509–512. doi: 10.3758/BF03329868

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: covariation assessment and Bayesian inference. *Cogn. Psychol.* 26, 2009–2239. doi: 10.1006/cogp.1994.1007

Nilsson, H. (2008). Exploring the conjunction fallacy within a category learning framework. *J. Behav. Decis. Making* 21, 471–490. doi: 10.1002/bdm.615

Nilsson, H., and Andersson, P. (2010). Making the seemingly impossible appear possible: effects of conjunction fallacies in evaluations of bets on football games. *J. Econ. Psychol.* 31, 172–180. doi: 10.1016/j.joep.2009.07.003

Nilsson, H., Juslin, P., and Winman, A. (2014). *Heuristics Can Produce Surprisingly Rational Probability Estimates: Comments on Costello and Watts (2014)*. Department of Psychology, Uppsala University, Uppsala, Sweden.

Nilsson, H., Rieskamp, J., and Jenny, M. A. (2013). Exploring the overestimation of conjunctive probabilities. *Front. Psychol.* 4:101. doi: 10.3389/fpsyg.2013.00101

Nilsson, H., Winman, A., Juslin, P., and Hansson, G. (2009). Linda is not a bearded lady: configural weighting and adding as the cause of extension errors. *J. Exp. Psychol. Gen.* 138, 517–534. doi: 10.1037/a0017351

Reyna, V. F., and Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learn. Individ. Differ.* 18, 89–107. doi: 10.1016/j.lindif.2007.03.011

Schneider, W., and Shiffrin R. M. (1977). Controlled and automatic human information processing: 1. detection, search, and attention. *Psychol. Rev.* 84, 1–66. doi: 10.1037/0033-295X.84.1.1

Shanteau, J. C. (1972). Descriptive versus normative models of sequential inference judgments. *Exp. Psychol.* 93, 63–68. doi: 10.1037/h0032509

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

Tenenbaum, J. B., Kemp, C., Griffith, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 19 November 2014; accepted: 13 January 2015; published online: 04 February 2015.

Citation: Juslin P (2015) Controlled information integration and bayesian inference. Front. Psychol. ***6****:70. doi: 10.3389/fpsyg.2015.00070*

This article was submitted to Cognition, a section of the journal Frontiers in Psychology.

frontiers in Psychology

OPINION
published: 22 May 2015
doi: 10.3389/fpsyg.2015.00680

Basic understanding of posterior probability

Vittorio Girotto * *and Stefania Pighin*

Center for Experimental Research on Management and Economics, Department of Culture Project, University IUAV of Venice, Venice, Italy

Keywords: posterior probability, updating, number of chances, natural frequencies, intuitive judgments

Consider the following task

[Task A]

A prenatal test determines whether an unborn child has a chromosomal anomaly. A priori, namely, before undergoing the test, a pregnant woman has a 4% chance of having a child with the anomaly. If a woman has a child with the anomaly, there is a 75% chance that she has a positive test result. If she does not have a child with the anomaly, there is still a 12.5% chance that she has a positive test result. Emma, a pregnant woman, undergoes a prenatal test. The result is positive. What is the probability that she has a child with the anomaly?

To answer correctly, one has to integrate the prior probability that a woman has a child with the anomaly (i.e., the prevalence rate: 4%) with information about the test's statistical properties. On the basis of this information and the evidence that Emma tested positive, one can produce a correct posterior evaluation by computing the ratio:

Probability (Anomaly|Positive Test Result) = Probability ("Positive Test Result and Anomaly")/ Probability ("Positive Test Result").

To obtain the numerator, one has to combine the prevalence rate and the test's sensitivity rate (i.e., 4% × 75% = 3%). To obtain the denominator, one has to combine the complement of the prevalence rate and the false positive rate (i.e., 96% × 12.5% = 12%), and then add it to the initially obtained value (i.e., 3% + 12% = 15%). Very few respondents, including health-care professionals, produce the correct probability ratio (i.e., 3%/15% = 20%). Failures to solve tasks of this sort lead to pessimistic conclusions about naive probabilistic reasoning (e.g., Casscells et al., 1978). Subsequent studies, however, licensed more optimistic conclusions, showing that some versions of these tasks led to better performances. About half of the respondents succeed when reasoning with natural frequencies (e.g., "Three out of the 4 women who had a child with the anomaly had a positive test result") or numbers of chances (e.g., "In 3 out of the 4 chances of having a child with the anomaly the test result is positive"; see, respectively, Hoffrage and Gigerenzer, 1998; Girotto and Gonzalez, 2001). On the basis of these results, the current, common account is that posterior probability reasoning improves in versions that allow respondents to both rely on an appropriate representation of subsets of countable elements (e.g., observations, tokens), and to easily associate posterior evidence with one of these subsets (Barbey and Sloman, 2007).

A generally unnoticed aspect of the results mentioned above is that they concern educated respondents, like undergraduates and physicians, and that only about half of these respondents benefit from the simplified versions of the tasks. Even more unnoticed is the fact that respondents sampled from the general public do not benefit at all from these versions. Indeed, in samples of pregnant women, many of whom had a high school level of education or less, almost all respondents failed to compute the correct probability ratio, even if they had to reason about natural frequencies (Bramwell et al., 2006) or numbers of cases (Pighin et al., 2015). In other words, they failed tasks that, in principle, should have activated the appropriate set representation. Their failure is striking because, unlike the participants of previous studies who had to reason about hypothetical

OPEN ACCESS

Edited by:
David R. Mandel, Defence Research and Development Canada, Toronto Research Centre, Canada

Reviewed by:
Ulrich Hoffrage, University of Lausanne, Switzerland
Stephanie Denison, University of Waterloo, Canada

***Correspondence:**
Vittorio Girotto, vgirotto@iuav.it

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *10 February 2015*
Accepted: *09 May 2015*
Published: *22 May 2015*

Citation:
Girotto V and Pighin S (2015) Basic understanding of posterior probability. Front. Psychol. 6:680. doi: 10.3389/fpsyg.2015.00680

scenarios, these women reasoned about realistic prenatal test results, and were personally interested in understanding them correctly.

In sum, contrary to the common account, naive respondents do not perform well on tasks devised to improve their understanding of posterior probability. These tasks mimic everyday problems, like calculating the post-test probability of diseases. However, they are unlikely to be the best tools to investigate whether naive respondents possess a basic intuition of posterior probability, and whether they are able to update their evaluations in the light of new evidence (Girotto and Gonzalez, 2007). Indeed, these tasks do not require respondents to revise any initial judgment (Girotto and Gonzalez, 2008; Mandel, 2014). Rather, they simply ask for only one judgment on the basis of various pieces of evidence (e.g., the prevalence rate, the result of the test and its statistical properties). Moreover, these verbal tasks convey numerical information by means of symbols and require an explicit numerical evaluation. Therefore, they can be employed only with literate respondents who have acquired a numerical symbolic system. Producing an explicit numerical estimation in numbers or words, however, is not the only way in which individuals may assess chance. Consider the following task:

[Task B]

Respondents are presented with a box containing five red chips (four round and one square) and three green chips (all square). The experimenter says, "I will take one chip out of the box without looking inside. Do you think that I will get a red or a green chip?"

Unlike Task A, and other verbal tasks used in adult Western literature, Task B does not convey probabilities by means of numerical symbols, and does not require respondents to produce an explicit numerical evaluation. Rather, it presents a set of tokens, and asks for a qualitative judgment or choice between two outcomes that may occur by taking one token out of the set at random (i.e., drawing a red vs. a green chip). To produce a suitable answer, respondents can reason extensionally, by considering and comparing the ways in which the outcomes may occur (Johnson-Laird et al., 1999). Accordingly, respondents will predict the occurrence of the outcome that may be produced in more ways (i.e., drawing a red chip). Numerate respondents could make a precise enumeration of the chances favoring each outcome (e.g., "There are 5 chances of drawing a red chip vs. 3 chances of drawing a green chip"). On this basis, they could even produce an explicit and correct absolute evaluation (e.g., "There are 5 chances out of 8 of drawing a red chip"). Of course, non-numerate respondents could not do so. However, the ability to make approximate comparisons of quantities emerges before (e.g., Barth et al., 2005) and without schooling (e.g., Pica et al., 2004). Therefore, even individuals who lack any formal numerical knowledge should produce suitable predictions in simple tasks like Task B. Indeed, both Western 5-year-olds (e.g., Davies, 1965; Girotto and Gonzalez, 2008) and preliterate Mayan adults (Fontanari et al., 2014) answer "red," that is, they choose the more likely outcome, and they do so even when they have to consider large sets of tokens. In sum, non-numerate individuals are able to compare the chances of two competing outcomes, without being able to express them numerically, and without necessarily making an explicit and precise counting of the number of chances favoring each of them.

Notably, these individuals also revise their evaluations on the basis of a new piece of evidence:

[Task B']

Upon the completion of Task B, the experimenter say, "I have taken one chip out of the box. I have it in my hand and I feel that it is square. Do you think that I got a red or a green chip?

To choose the more likely outcome ("green"), respondents should focus on the subset of possibilities compatible with the evidence (the four squares). Five-year-olds do so, updating their initial judgments and choices suitably (Girotto and Gonzalez, 2008/Studies 1 and 2). They succeed even in tasks that imply more complex combinations of prior and posterior information (Bonawitz et al., 2013), or reasoning about a single, non-repeatable event produced by an intentional agent (Girotto and Gonzalez, 2008/Study 3). Fontanari et al. (2014) have extended these results by presenting preliterate Mayan adults with the same sort of tasks. Despite their lack of any sort of formal education, these respondents performed like Western controls, revising their initial choices in the light of new evidence. Finally, measures of looking times suggest that even preverbal infants form rational expectations about uncertain events by integrating different sources of information in a coherent way (Teglas et al., 2011). Together, these findings corroborate the view that, along with the application of non-extensional heuristics (Tversky and Kahneman, 1974), naive reasoning about probabilities often relies on extensional procedures: respondents infer the probability of an event from the various ways in which it could occur (Johnson-Laird et al., 1999).

Two notes are in order about the tasks that have documented the existence of an early understanding of prior and posterior probability (e.g., Task B and B'). First, these tasks are not natural frequency tasks. Indeed, they do not convey natural frequency information and do not ask for a frequency prediction. The following one is an example of a proper natural frequency task:

[Task C]

The experimenter says, "This box contains some chips. You do not know their colors. You observe me drawing a chip at random from the box, and replacing it in the box 8 times. My sample shows 5 red and 3 green chips. I'll draw a chip at random 8 more times. Do you think that the new sample will show more red or more green chips?"

Task C is apparently similar to Task B. In both cases, one can answer by considering sets of countable elements (i.e., prior possibilities and actual frequencies, respectively), and by making a similar comparison (i.e., 5 red chips vs. 3 green chips, and 5 draws of a red chip vs. 3 draws of a green chip, respectively). The two answers, however, cannot be assimilated. In Task B, one reasons about a set of prior possibilities before making any actual experience. In Task C, one reasons about a set of observations gathered through a "natural sampling," which is "the process of encountering instances in a population sequentially. The outcome of natural sampling is natural frequencies" (Gigerenzer and Hoffrage, 1999, p. 425).

Second, tasks that do not ask for an explicit numerical evaluation, including those that imply reasoning about few possibilities, do not guarantee correct performance neither in children nor in adults (Nickerson, 1996; Johnson-Laird et al., 1999). Consider, for example, Task B. Young children succeed in it, basing their answer on prior possibilities (e.g., "You will get a red chip because there are more red than green chips"). However, if one transforms Task B into a frequency-like task, they fail. In other words, if one makes a series of random draws from the same box, and asks young children to make a prediction for each of them, they tend to use erroneous strategies like "Predict the color that was not predicted in the previous trial" (Brainerd, 1981; Teglas et al., 2007/Studies 3 and 4). It should be noted that even literate adults make erroneous predictions in situations in which they have to extract frequencies from actual observations rather than to process numerical symbols. For example, they fail versions of Task A in which they are presented with a series of medical records, each representing a patient, his/her health condition and the presence/absence of a given symptom (e.g., Gluck and Bower, 1988). Along with the finding that young children can reason correctly about events before experiencing their actual frequency, the finding that literate adults err in experience-based reasoning tasks is difficult to explain following the hypothesis that the human mind is "developmentally and evolutionary prepared to handle natural frequencies" (Gigerenzer and Hoffrage, 1999, p. 430).

In conclusion, even literate adults have difficulties in producing correct posterior evaluations. They appear to be unable to combine prior information and new evidence in a normative way in tasks whose solution depends on the combination of numerical values, including tasks that have been devised to improve posterior probability reasoning. However, recent studies have shown that even young children and preliterate adults can succeed in tasks whose solution depend on a simple comparison of possibilities. In sum, naive individuals possess correct intuitions of prior and posterior probabilities, and such intuitions emerge early in the course of development and regardless of culture and education.

Acknowledgments

The authors thank Ulrich Hoffrage, David Mandel and Stephanie Denison for their comments on a previous version of this paper, which was supported by grants from Swiss&Global—Ca' Foscari Foundation and the Italian Ministry of Scientific Research (PRIN grant 2010-RP5RNM).

References

Barbey, A. K., and Sloman, S. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–297. doi: 10.1017/S0140525X07001653

Barth, H., Le Mont, K., Lipton, J., and Spelke, E. S. (2005). Abstract number and arithmetic in preschool children. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14116–14121. doi: 10.1073/pnas.0505512102

Bonawitz, E., Denison, S., Griffith, T. L., and Gopnick, A. (2013). Rational variability in children's causal inferences: the sampling hypothesis. *Cognition* 126, 285–300. doi: 10.1016/j.cognition.2012.10.010

Brainerd, C. J. (1981). Working memory and the developmental analysis of probability judgment. *Psychol. Rev.* 88, 463–502. doi: 10.1037/0033-295X.88.6.463

Bramwell, R., West, H., and Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: experimental study. *Br. Med. J.* 333, 284–289. doi: 10.1136/bmj.38884.663102.AE

Casscells, W., Schoenberger, A., and Graboys, T. (1978). Interpretation by physicians of clinical laboratory results. *N. Engl. J. Med.* 299, 999–1000. doi: 10.1056/NEJM197811022991808

Davies, C. M. (1965). Development of the probability concept in children. *Child Dev.* 36, 779–788. doi: 10.2307/1126923

Fontanari, L., Gonzalez, M., Vallortigara, G., and Girotto, V. (2014). Probabilistic cognition in two indigenous maya groups. *Proc. Natl. Acad. Sci. U.S.A.* 111, 17075–17080. doi: 10.1073/pnas.1410583111

Gigerenzer, G., and Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: a reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychol. Rev.* 106, 425–430. doi: 10.1037/0033-295X.106.2.425

Girotto, V., and Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition* 78, 247–276. doi: 10.1016/S0010-0277(00)00133-5

Girotto, V., and Gonzalez, M. (2007). How to elicit sound probabilistic reasoning: beyond word problems. *Behav. Brain Sci.* 30, 268. doi: 10.1017/S0140525X07001768

Girotto, V., and Gonzalez, M. (2008). Children's understanding of posterior probability. *Cognition* 106, 325–344. doi: 10.1016/j.cognition.2007.02.005

Gluck, M. A., and Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *J. Exp. Psychol. Gen.* 117, 227–247. doi: 10.1037/0096-3445.117.3.227

Hoffrage, U., and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Acad. Med.* 73, 538–540. doi: 10.1097/00001888-199805000-00024

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M., and Caverni, J. P. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychol. Rev.* 106, 62–88.

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Front. Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

Nickerson, R. S. (1996). Ambiguities and unstated assumptions in probabilistic reasoning. *Psychol. Bull.* 120, 410–430. doi: 10.1037/0033-2909.120.3.410

Pica, P., Lerner, C., Izard, V., and Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science* 306, 499–503. doi: 10.1126/science.1102085

Pighin, S., Gonzalez, M., Savadori, L., and Girotto, V. (2015). Improving public interpretation of probabilistic test results: distributive evaluations. *Med. Decis. Making* 35, 12–15. doi: 10.1177/0272989X14536268

Teglas, E., Girotto, V., Gonzalez, M., and Bonatti, L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19156–19159. doi: 10.1073/pnas.0700271104

Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., and Bonatti, L. (2011). Pure reasoning in 12-months-old infants as probabilistic inference. *Science* 332, 1054–1059. doi: 10.1126/science.1196404

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

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

OPINION
published: 02 June 2015
doi: 10.3389/fpsyg.2015.00712

Beyond getting the numbers right: what does it mean to be a "successful" Bayesian reasoner?

Gaëlle Vallée-Tourangeau[1]*, *Miroslav Sirota*[1], *Marie Juanchich*[2] and *Frédéric Vallée-Tourangeau*[1]

[1] Department of Psychology, Kingston University London, Kingston upon Thames, UK, [2] Department of Management, Kingston University London, Kingston upon Thames, UK

Keywords: Bayesian inference, insight problem solving, judgment and decision-making, performance, cognitive processes

Edited by:
David R. Mandel,
Defence R&D Canada, Canada

Reviewed by:
Simon John McNair,
Leeds University Business School, UK
Peter Juslin,
Uppsala University, Sweden

***Correspondence:**
Gaëlle Vallée-Tourangeau,
g.vallee-tourangeau@kingston.ac.uk

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *09 February 2015*
Accepted: *13 May 2015*
Published: *02 June 2015*

Citation:
Vallée-Tourangeau G, Sirota M, Juanchich M and Vallée-Tourangeau F (2015) Beyond getting the numbers right: what does it mean to be a "successful" Bayesian reasoner? Front. Psychol. 6:712. doi: 10.3389/fpsyg.2015.00712

Price (in Bayes, 1958) introduced Bayes's theorem as a precise and accurate method for measuring the strength of an inductive argument. He contrasted Bayesian reasoning with common sense, which, he argued, is imbued with vagueness and often erroneous. Nearly two centuries later, Price's claim was put to the test by psychologists who examined how people revise their opinions in light of new evidence (e.g., Phillips and Edwards, 1966; Kahneman and Tversky, 1973). For the past four decades, scholars have debated whether common sense can or cannot approximate Bayesian reasoning.

Contrary to Price's claim, earlier studies using a bookbags-and-poker-chips paradigm found that people did follow Bayesian prescriptions to revise judgments although their numerical answers were conservative: the psychological impact of new evidence on one's belief was less pronounced than warranted (Edwards, 1968). A paradigm shift ensued with the advent of the heuristic-and-biases programme of research (Kahneman et al., 1982). Scholars started to use vignette studies modeled after the so-called "textbook paradigm" or the "social-judgment paradigm" (Bar-Hillel, 1983). This also led to an about-turn in the portrayal of people's ability to revise their judgments accurately. Vignette studies did not showcase mere conservatism, they elicited biased judgments which were often in blatant contradiction with Bayesian prescriptions. This bleak picture of people's ability to form Bayesian judgments was once more overturned in the mid-nineties when researchers demonstrated that natural frequency formats could lead to a fourfold improvement in performance rates (Gigerenzer and Hoffrage, 1995; Cosmides and Tooby, 1996). This finding once more shifted the point of scholarly contention as scholars started to debate whether the improvement observed arises from the use of natural frequencies in and by itself or from the more effective "nested representation" that this information format elicits (Sirota et al., 2015).

Throughout this (admittedly) short history of the psychological study of Bayesian reasoning, Bayesian performance has most commonly been defined, explicitly or implicitly, as the ability to generate the "accurate" value for the posterior probability $p(H|D)$, or the probability that a hypothesis H is true, given a new piece of evidence D, based on the values of $p(H)$, $p(\text{not-}H)$, $p(D|H)$ and $p(D|\text{not-}H)$ where $p(H)$ denotes the *a priori* probability that H is true and $p(\text{not-}H)$, the *a priori* probability that its alternative, not-H is true (which may or may not be equated with the base rates; Mandel, 2014); $p(D|H)$ denotes the probability of observing D when we know H to be true; and, finally, $p(D|\text{not-}H)$ denotes the probability of observing D when the alternative hypothesis, not-H is true.

This approach to performance assessment—comparing a normative numerical value to a subjective probability estimate—informs *what* is computed (a Bayesian answer, based on a correct number or a correct algorithm) and enables researchers to assess Bayesian *performance*. Efforts to improve Bayesian performance have focused on modifying environmental characteristics such

as the probabilistic information format (e.g., Gigerenzer and Hoffrage, 1995). But performance arises from the coupling of the task environment and the cognitive processes applied to the task at hand. Fostering better Bayesian performance can also involve a better understanding of Bayesian *reasoning*, that is, *how* the subjective estimate is actually computed (e.g., see Sirota et al., 2014).

Adopting a reasoning-based focus also sheds light on differences between the three classic paradigms mentioned above that would otherwise remain concealed. While any of these paradigms may be used interchangeably to assess Bayesian *performance*, whether they all involve the same type of Bayesian *reasoning* is debatable. This is not a trivial distinction: if different paradigms invoke different reasoning processes, what works for improving Bayesian performance will be contingent on the particular research paradigm adopted to study performance. In the remainder of this essay, we show that a focus on performance (where participants' probability judgments are compared to Bayesian normative values or algorithms) obscures the fact there are more than one way to engage in Bayesian reasoning. Our analysis suggests three criteria against which the quality of Bayesian inferences may be assessed: an *accuracy criterion* (did participants compute the normative value? Did they apply the correct algorithm?), an *adequacy criterion* (did participants appropriately revise their initial judgment?), and a *restructuring criterion* (did participants successfully restructure their initial representation of the problem to achieve the goal state?).

The typical bookbags task involves two urns with symmetrical assortment of marbles—e.g., a "black urn" with 600 black and 400 white marbles, and a "white urn" with 400 white and 600 black marbles (Peterson et al., 1965). An experimenter selects one urn at random and hides it in an opaque box from which he then draws several samples of marbles. After observing each sample, participants are asked to revise the probability that the sample originates from one urn by moving a slider along a bar displaying 100 marks. The length of the bar's left section represents the probability that the marbles had been drawn from the black urn. Participants' output judgments can be compared with the Bayesian norm. This involves computing $p(D|H)$ and $p(D|$ not-$H)$, the probabilities of observing the sample D if it were obtained from the black urn and the white urn, respectively. Even when participants are informed about the exact ratio of marbles in each urn, it is implausible to assume that they engage in such explicit numerical computations to revise their judgment. Instead, belief revision is more likely to arise from intuitive thinking processes involving an assessment of the perceptual similarity between the sample and the urn (e.g., see Read and Grushka-Cockayne, 2011). In such a context, interventions on feedback and learning from experience are more likely to improve Bayesian reasoning than manipulations of information format, for example.

Social-judgment studies of Bayesian reasoning (e.g., Kahneman and Tversky, 1973) include social scenarios and subjective probabilities implied by thumbnail verbal descriptions instead of countable numerical information. Typically, social-judgment tasks involve the assessment of the posterior probability that an individual belongs to a target category (e.g., engineer), based on both a short verbal description of the individual's social attributes (e.g., "spends most of his free time on his many hobbies which include home carpentry, sailing and mathematical puzzles" Kahneman and Tversky, 1973, p. 241) and the numerical base rate of the target category and an alternative category (e.g., 30 engineers and 70 lawyers). So while social-judgment tasks provide precise information about the base rates, the numerical values of the likelihood probabilities $p(D|H)$ and $p(D|$ not-$H)$ of the descriptions are neither presented to, nor elicited from the participants. By comparing subjective posterior probability judgments made in this instance with judgments made for reversed base-rate distributions (e.g., 70 engineers and 30 lawyers), it is possible to evaluate the extent to which judgments are aligned with Bayesian prescriptions just as with the bookbags paradigm. Once again, however, these judgments are unlikely to arise from explicit numerical computations akin to those required to compute the Bayesian benchmark criterion since this would require that participants spontaneously generate a numerical value for $p(D|H)$ and $p(D|$ not-$H)$. In fact, the actual origin of the estimate produced by participants in Social-judgment tasks is unclear. The attribute-substitution account (Kahneman and Frederick, 2002) theorizes that participants use a heuristic attribute (e.g., the extent to which the individual described is similar to a typical engineer) as a substitute for the target attribute (e.g., the probability that the individual is an engineer, given his description) in their assessment. This account, however, does not explain *how* people may compute the similarity index between the verbal description of an individual instance and a typical instance. Dougherty et al.'s (1999) MINERVA Decision-Making (MDM) model proposes that judgments are based on less than perfect memory retrieval of observations frequencies. The predictive value of the MDM model is established by comparing averaged simulated outputs with Bayesian computations and demonstrating that the simulations derived from the model are consistent with actual judgments observed in Social-judgment studies. This model is underpinned by two assumptions: first, that social judgments have a frequentist origin, and second that all individuals rely on the same memory-based process to compute their judgment. Both assumptions have yet to be tested empirically. In sum, more research is needed before the cognitive processes that yield such judgment methods in Bayesian reasoning can be firmly established. In this respect, representational theories of subjective probability such as Mandel's (2008) representational and assessment processes account may prove fruitful.

The last, and perhaps most prevalent, paradigm is the so-called textbook one. In this paradigm, participants are presented with explicit numerical values for all the components required for computing the posterior probability $p(H|D)$, namely $p(H)$, $p(D|H)$ and $p(D|$ not-$H)$ as in, for example, the mammography problem (Gigerenzer and Hoffrage, 1995). Once again, performance may be assessed in the same way it is assessed in bookbags tasks or in social judgment tasks: by comparing participants' judgment to the Bayesian criterion. The reasoning processes, which lead to the final judgment, however, are unlikely to be based on assessments of perceptual similarities (as in bookbags tasks) or memory retrieval of observed frequencies (as in social judgment tasks). Instead, textbook tasks require

participants to reach a goal state (the posterior probability value) based on an initial state presenting the values of the base rate, hit rate and false alarm probabilities. In other words, textbook tasks require participants to apply operators to move from an initial state (the problem presentation) to a series of different states until the final goal state is reached. These tasks do not require an intuitive judgment of a probability value, they require analysis and problem-solving skills. As such, problem-solving theory can shed new light on the processes that underpin Bayesian reasoning in textbook problems.

Problem-solving theorists often distinguish between routine and non-routine problems (e.g., see Mayer, 1995). Routine problems involve the application of a known procedure to be solved. For example, 2 + 2 is a routine problem for anyone who has been taught a procedure for adding single digits. Applying the known procedure involves reproductive thinking; once the procedure is known, problem solvers can apply it again to solve similar problems. By contrast, when problem-solvers face non-routine problems, they do not possess a pre-existing solution procedure; they must engage in productive thinking and generate a novel solution to reach the goal state. Textbook problems presented to naive participants, that is participants who have not learnt to apply the Bayesian procedure to compute $p(H|D)$, are difficult non-routine problems. Problem solvers may have some operators which they can apply (like adding values or multiplying them) but they have no means to gauge their progress or assess the validity of their final answer. This suggests that a possible way forward to better understand how participants may succeed in textbook tasks would be to consider those tasks as insight problems. From a set theoretic perspective, the prior probability $p(H)$ corresponds to the proportion of the sample space S that is occupied by *H*. The occurrence of the outcome *d* reduces the sample space to the event *D* because the elements outside *D* are no longer possible outcomes. Consequently, the probability of *H given D* is the probability of *H* given the reduced sample space *D*. This analysis suggests that Bayesian performance in textbook problems demands that reasoners restructure their initial representation from the sample space S defined by the union of subsets *H* and *not-H* that both include *d* elements to the subset *D* that includes *h* and *not-h* elements.

To sum up, in this essay, we argued for a distinction between Bayesian performance and Bayesian reasoning. Whereas Bayesian performance can be assessed through a variety of paradigms, a focus on performance obscures the fact there are more than one way to engage in Bayesian reasoning: people may reason appropriately but perform poorly, thus committing what is known as an "error of application" (Kahneman and Tversky, 1982). Conversely, they may adopt an inappropriate line of reasoning (thus committing an "error of comprehension," Kahneman and Tversky, 1982) but nevertheless produce an accurate judgment. Our analysis suggests three criteria against which the quality of Bayesian inferences may be assessed: an accuracy criterion, an adequacy criterion, and a restructuring criterion. Whereas the accuracy criterion is applicable in all three paradigms, the adequacy criterion is better suited to bookbags tasks because they require participants to revise an initial judgment or the social-judgment tasks because they ask participants to provide a subjective estimate that weighs numerical-explicit and subjective-implicit information. Likewise, the restructuring criterion is better suited to textbook tasks as these tasks require participants to navigate through a problem space. Each criterion also points to different strategies for improving the quality of Bayesian inferences. The accuracy criterion favors analytical accounts where reasoning is defined as the step-by-step transformation of explicit numerical quantities and facilitation results from easing the cognitive cost of carrying out these computations. The adequacy criterion favors associative accounts where reasoning is defined as belief updating and facilitation results from the better calibration of the subjective weights attributed to different inputs. Finally, the restructuring criterion favors representational accounts where reasoning is defined as navigating through a problem space and facilitation results from the clarification of the representational structure of the problem. In other words, better understanding how people arrive at their answers in the different paradigms may prove a fruitful way forward to uncover the keys to further improve the quality of naive Bayesian inferences.

References

Bar-Hillel, M. (1983). "The base rate fallacy controversy," in *Decision Making Under Uncertainty: Cognitive Decision Research, Social Interaction, Development and Epistemology*, ed R. W. Scholz (Amsterdam: Elsevier Science), 39–61.

Bayes, T. (1958). Essay towards solving a problem in the doctrine of chances. *Biometrika* 45, 293–315.

Cosmides, L., and Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* 58, 1–73.

Dougherty, M. R. P., Gettys, C. F., and Ogden, E. E. (1999). MINERVA-DM: a memory processes model for judgments of likelihood. *Psychol. Rev.* 106, 180–209.

Edwards, W. (1968). "Conservatism in human information processing," in *Formal Representation of Human Judgment*, ed B. Kleinmuntz (New York, NY: Wiley), 17–52.

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704.

Kahneman, D., and Frederick, S. (2002). "Representativeness revisited: attribute substitution in intuitive judgment," in *Heuristics and Biases: The Psychology of Intuitive Judgment*, eds T. Gilovich, D. Griffin, and D. Kahneman (Cambridge, UK: Cambridge University Press), 49–81.

Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychol. Rev.* 80, 237–251.

Kahneman, D., Slovic, P., and Tversky, A. (eds.). (1982). *Judgment Under Uncertainty: Heuristic and Biases*. Cambridge, UK: Cambridge University Press.

Kahneman, D., and Tversky, A. (1982). On the study of statistical intuitions. *Cognition* 11, 123–141.

Mayer, R. E. (1995). "The search for insight: grappling with Gestalt psychology's unanswered questions," in *The Nature of Insight*, eds R. J. Sternberg and J. E. Davidson (Cambridge, MA: MIT Press), 3–32.

Mandel, D. R. (2008). Violations of coherence in subjective probability: a representational and assessment processes account. *Cognition* 106, 130–156. doi: 10.1016/j.cognition.2007.01.001

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Front. Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

Peterson, C. R., Schneider, R. J., and Miller, A. J. (1965). Sample size and the revision of subjective probability. *J. Exp. Psychol.* 69, 522–527.

Phillips, L. D., and Edwards, W. (1966). Conservatism in a simple probability model inference task. *J. Exp. Psychol.* 72, 346–354.

Read, D., and Grushka-Cockayne, Y. (2011). The similarity heuristic. *J. Behav. Decision Making* 24, 23–46. doi: 10.1002/bdm.679

Sirota, M., Kostovièová, L., and Juanchich, M. (2014). The effect of iconicity of visual displays on statistical reasoning: evidence in favor of the null hypothesis. *Psychon. Bull. Rev.* 21, 961–968. doi: 10.3758/s13423-013-0555-4

Sirota, M., Kostovièová, L., and Vallée-Tourangeau, F. (2015). How to train your Bayesian: a problem-representation transfer rather than a format-representation shift explains training effects. *Q. J. Exp. Psychol.* 68, 1–9. doi: 10.1080/17470218.2014.972420

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in PSYCHOLOGY

OPINION ARTICLE
published: 06 November 2014
doi: 10.3389/fpsyg.2014.01272

Communicating risk in prenatal screening: the consequences of Bayesian misapprehension

Gorka Navarrete[1]*, Rut Correia[2] and Dan Froimovitch[3]

[1] Laboratory of Cognitive and Social Neuroscience, Department of Psychology, Universidad Diego Portales, UDP-INECO Foundation Core on Neuroscience, Santiago, Chile
[2] Faculty of Education, Universidad Diego Portales, Santiago, Chile
[3] Department of Physiology, University of Toronto, Toronto, ON, Canada
*Correspondence: gorkang@gmail.com

Edited by:
David R. Mandel, Defence Research and Development Canada, Canada

Reviewed by:
Miroslav Sirota, King's College London, UK
Simon John McNair, Leeds University Business School, UK

Keywords: Bayesian reasoning, prenatal screening, health policies, risk communication, massive screening

At some point during pregnancy women are typically encouraged to undergo a screening test in order to estimate the likelihood of fetal chromosomal aberrations. While timelines vary, the majority of pregnant women are screened within their first trimester (De Graaf et al., 2002). In the event of a positive test result, an invasive diagnostic assessment is usually recommended, namely amniocentesis or chorionic villus sampling (CVS). The combined test, widely considered to be the most feasible and effective screening procedure, involves an integrated assessment of: maternal age, fetal Nuchal Translucency (NT), maternal serum pregnancy-associated plasma protein A (PAPP-A), and free β human chorionic gonadotropin (β-hCG). This assay is most reliable when performed nearest to the 11th week of gestation (Malone et al., 2005), at which its detection rate and false positive rate for trisomy 21, in optimal conditions, are approximately 95 and 5%, respectively (Nicolaides, 2004). A variety of competing screening techniques are available in the first trimester, and though we focus on the combined test in our example below, the point raised in this article applies to each of them.

A first-trimester screening assay carrying a relatively low false-positive rate might seem a reasonable option for women already considered to be at low risk—the vast majority of the pregnant population. Following such prenatal screening for trisomy 21, most women who test positive for high risk proceed with invasive diagnostic testing. This decision to proceed with invasive testing is typically based on the presence of any evidence of increased risk brought to light by the precursory screening test (Nicolaides, 2004). It is important to note, however, that the proportion of those who advance to invasive diagnostic testing is virtually identical to the false-positive rate of initial screening (Nicolaides, 2004).

Applying trisomy 21 as an example (see **Figure 1** for a graphical representation of the numbers), the pregnant women who receive a false positive score in their first-trimester screening (∼5%) would subsequently undergo a supplementary invasive diagnostic procedure, such as amniocentesis or CVS. This implies that out of every 100,000 pregnant women initially screened, roughly 5100 test positive, out of which ∼5000 cases are actually false positives. The follow-up diagnostic tests are associated with serious procedure-related health risks, including a ∼1% increased chance of miscarriage (see Mujezinovic and Alfirevic, 2007 for a systematic review; also, a recent nation-wide 11-year longitudinal study in Denmark established an increased chance of miscarriage of 1.4% and 1.9% linked to amniocentesis and CVS respectively, with CVS growing in its predominance worldwide; Tabor et al., 2009). Thus, at least 50 of the above ∼5000 false-positive cases that involve normal fetuses ultimately result in diagnostic procedure-induced miscarriage. Of course with either a higher false-positive rate or a lower disease prevalence, those numbers worsen.

Discerning the trustworthiness of a given positive result in a screening test warrants calculating (typically from the information provided in the respective consent form) the test's positive predictive value (PPV; in this case the proportion of Down syndrome cases relative to the total amount of positive results). This requires knowledge of the base incidence rate of the congenital defect of interest, and the sensitivity and false-positive rate of the test. Computation and proper interpretation of this index, however, is often obscured by the complexity of Bayesian reasoning involved. This, among other factors, may underlie the well-known inadequacy of current procedures intended to achieve informed consent (Green et al., 2004). For 30-year-old pregnant women, the prevalence of Down syndrome is roughly 1 out of every 800 fetuses (Nicolaides, 2004; this statistic varies with maternal age and time-point during pregnancy). In a sample of 100,000 pregnant women of the general population, therefore, around 125 of them would be expected to carry a fetus with the condition. Given the relatively high sensitivity of the screening assay (95% in optimal conditions), a majority of those fetuses are eventually correctly diagnosed with Down Syndrome (∼119 out of 125). But when we merge this information with the said ∼5000 false positives, we see that 119 positive results in the combined test faithfully reveal trisomy 21, out of a total 5113 (119 + 4994) positive results. Hence, the PPV of the combine test in a screening context nears 2% (119/5113). In other

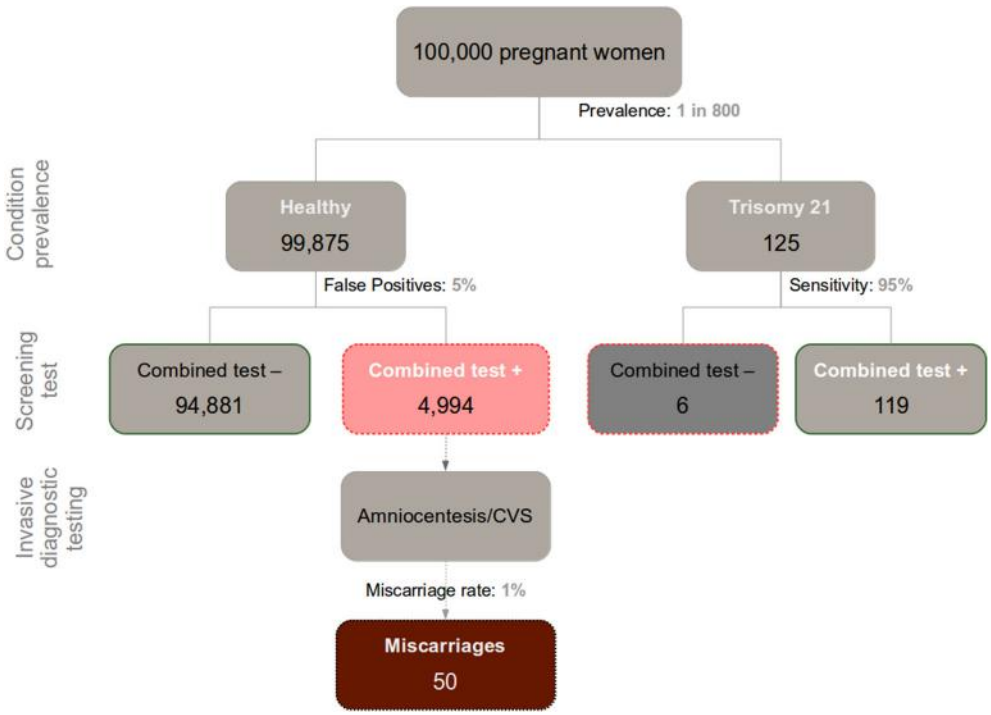

FIGURE 1 | Chart depicting the relationship between incidence of Down Syndrome (Trisomy 21), false positives in prenatal screening, and miscarriages caused by the recommended follow-up diagnostic assessment (Amniocentesis/CVS) in a sample of 100,000 pregnant women.

words, there is a 2% chance of actually carrying a fetus with trisomy 21 after testing positive in a screening combined test. This information—essential to an educated decision on the matter—is usually overlooked by practitioners, and generally absent from medical consent forms.

In recent decades, our ineptitude for making sense of Bayesian information has been the subject of extensive study (for a review see Barbey and Sloman, 2007). It is widely recognized that humans struggle in dealing with Bayesian problems presented in terms of normalized probabilities (i.e., relative probabilities or percentages) or in cases of vague information structure (Barbey and Sloman, 2007). A substantial portion of the research on this topic has been done within the scope of medicine and epidemiology, wherein Bayesian inference pervades disease detection and characterization. It is well known that even medical practitioners struggle to interpret such information (Gigerenzer et al., 2007; but see Pighin et al., 2014 for a more optimistic outlook). The issue saliently manifests in the prevailing appeal of massive screening programs to the general public, policy-makers, and physicians alike. This appeal—mainly due to the perceived advantages of early diagnosis—fails to be balanced by sufficient consideration of the high propensity for false alarms and over-diagnosis. The theoretic difficulties that most primary care physicians, for instance, seem to encounter with this type of information (e.g., cancer screening statistics) disposes them to a disproportionate veneration for the potential benefits of disease screening, as they drastically underrate the seriousness of relevant risks.

Gigerenzer et al. have advised on the pernicious use of massive screenings with respect to prostate cancer, HIV infection, etc. (Gigerenzer et al., 2007). False positives can be highly problematic in their ensuing psychosocial turmoil, and with respect to iatrogenic complications and economic costs associated with unnecessary clinical intervention. Moreover the problems, as we have seen above, don't stop at this. Medical knowledge ought to be conveyed lucidly, in a manner that facilitates informed decision-making, specifically accounting for the common cognitive challenges and inter-individual variation observed in probability literacy (Johnson and Tubau, 2013; Lesage et al., 2013; Låg et al., 2014; Sirota et al., 2014a). With respect to clinical screening data, sufficient understanding of the numbers not only entails being in a position to competently evaluate pertinent risks; it further entails being enabled to recognize the possibility that even tests carrying low false-positive rates may simply be inadequate for detecting low-prevalence diseases, particularly in massive-screening settings.

There is growing convergence in cognitive psychology regarding the chief factors that mediate computation of Bayesian reasoning problems. Furthermore some practical improvements in the communication of statistical information have been proposed (while focus on evolutionary underpinnings of these issues appears to have taken a back seat in the literature (Barbey and Sloman, 2007; Navarrete and Santamaría, 2011). With respect to understanding Bayesian problems, apart from intrinsic differences across individuals, in cognitive resources (Lesage et al., 2013; Låg et al., 2014; Sirota et al., 2014a) or numeracy skill (Hill and Brase, 2012; Johnson

and Tubau, 2013; Låg et al., 2014), several other factors that pertain to informational presentation *per se* have been deemed relevant to reasoning performance. These include (but are not limited to): problem structure (Barbey and Sloman, 2007; Lesage et al., 2013; Sirota et al., 2014a), the availability of a causal framework (Krynski and Tenenbaum, 2007), representational format (Hoffrage et al., 2002), and reference class (Fiedler et al., 2000; Lesage et al., 2013). Over and above intellectual aptitude, the very manner in which a problem's terms are conveyed to the subject is arguably imperative to the normative Bayesian response.

The above theoretical advancements have translated into numerous helpful strategies for representing and communicating Bayesian information. Regarding medical risk problems, if a subject is provided with the relevant information comprising the standard menu (i.e., hit rate, false positive rate and prevalence; Gigerenzer and Hoffrage, 1995), the most effective way known to facilitate reasoning is to ensure that the problem's set structure is entirely clarified to the subject (Barbey and Sloman, 2007). Natural frequencies (Gigerenzer and Hoffrage, 1995), or more generally, absolute reference classes (Fiedler, 2000; Lesage et al., 2013) are widely considered instrumental to this end. Another important factor, admittedly difficult to disentangle conceptually from the previous one, is computational complexity (Gigerenzer and Hoffrage, 1995; Barbey and Sloman, 2007). Reducing a subject's need to carry out computations (even those of simple arithmetic operations) can substantially enhance reasoning performance. Moreover the use of iconic and interactive representations has been shown to improve performance accuracy (Brase, 2009; Tsai et al., 2011; Micallef et al., 2012; Sirota et al., 2014b). Finally, an increasingly important area of research in this regard pertains to the development of training-programs designed to improve patients' and physicians' comprehension and computation of Bayesian problems (Sedlmeier and Gigerenzer, 2001; Sirota et al., 2014c).

There is a persistent need for advancing research concerning efficacious communication of Bayesian information, such that it can be comprehended by as many individuals as possible—most urgently, those who intervene in health care decision making, such as clinicians and policy-makers. Wide-scale disease screenings hold both advantages and drawbacks (Gigerenzer et al., 2007), and a clear cognizance of their performance characteristics and the numbers underlying them is crucial to the state of public health and safety. At the moment, however, sufficient understanding of them is strikingly scarce, and with each passing year an unacceptable number of prospective parents are pressed to carry out a critical decision of potentially daunting consequences, without adequate knowledge of the important risks. And, of course, the quintessential challenges inherent to Bayesian reasoning are appreciable well beyond the domain of prenatal screening, posing egregious threats to the security and well-being of both the individual and the public.

ACKNOWLEDGMENTS

This research has been supported by the Semilla grants from the University Diego Portales (SEMILLA201418).

REFERENCES

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci*. 30, 241–254. doi: 10.1017/S0140525X07001653

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol*. 23, 369–381. doi: 10.1002/acp.1460

De Graaf, I. M., Tijmstra, T., Bleker, O. P., and van Lith, J. M. M. (2002). Womens' preference in Down syndrome screening. *Prenat. Diagn*. 22, 624–629. doi: 10.1002/pd.358

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychol. Rev*. 107, 659–676. doi: 10.1037/0033-295X.107.4.659

Fiedler, K., Brinkmann, B., Betsch, T., and Wild, B. (2000). A sampling approach to biases in conditional probability judgments: beyond base rate neglect and statistical format. *J. Exp. Psychol. Gen*. 129, 399–418. doi: 10.1037/0096-3445.129.3.399

Gigerenzer, G., Gaissmaier, W., Kurz-milcke, E., Schwartz, L. M., Woloshin, S., and Dartmouth, T. (2007). Helping doctors and patients make sense of health statistics. *Psychol. Sci. Public Interest* 8, 53. doi: 10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev*. 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Green, J. M., Hewison, J., Bekker, H. L., Bryant, L. D., and Cuckle, H. S. (2004). Psychosocial aspects of genetic screening of pregnant women and newborns: a systematic review. *Health Technol. Assess*. 8, 1–109.

Hill, W. T., and Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *Q. J. Exp. Psychol*. 65, 2343–2368. doi: 10.1080/17470218.2012.687004

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition* 84, 343–352. doi: 10.1016/S0010-0277(02)00050-1

Johnson, E. D., and Tubau, E. (2013). Words, numbers, and numeracy: diminishing individual differences in Bayesian reasoning. *Learn. Individ. Differ*. 28, 34–40. doi: 10.1016/j.lindif.2013.09.004

Krynski, T. R., and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *J. Exp. Psychol. Gen*. 136, 430–450. doi: 10.1037/0096-3445.136.3.430

Låg, T., Bauger, L., Lindberg, M., and Friborg, O. (2014). The role of numeracy and intelligence in health-risk estimation and medical data interpretation. *J. Behav. Decis. Mak*. 27, 95–108. doi: 10.1002/bdm.1788

Lesage, E., Navarrete, G., and De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: the role of general cognitive resources. *Think. Reason*. 19, 27–53. doi: 10.1080/13546783.2012.713177

Malone, F. D., Canick, J. A., Ball, R. H., Nyberg, D. A., Comstock, C. H., Bukowski, R., et al. (2005). First-trimester or second-trimester screening, or both, for Down's syndrome. *N. Engl. J. Med*. 353, 2001–2011. doi: 10.1056/NEJMoa043693

Micallef, L., Dragicevic, P., and Fekete, J. (2012). Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Trans. Vis. Comput. Graph*. 18, 2536–2545. doi: 10.1109/TVCG.2012.199

Mujezinovic, F., and Alfirevic, Z. (2007). Procedure-related complications of amniocentesis and chorionic villous sampling: a systematic review. *Obstet. Gynecol*. 110, 687–694. doi: 10.1097/01.AOG.0000278820.54029.e3

Navarrete, G., and Santamaría, C. (2011). Ecological rationality and evolution: the mind really works that way? *Front. Psychol*. 2:251. doi: 10.3389/fpsyg.2011.00251

Nicolaides, K. H. (2004). *The 11–13+6 Weeks Scan*. London: Fetal Medicine Foundation.

Pighin, S., Gonzalez, M., Savadori, L., and Girotto, V. (2014). Improving public interpretation of probabilistic test results: distributive evaluations. *Med. Decis. Mak*. doi: 10.1177/0272989X14536268. [Epub ahead of print].

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen*. 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Sirota, M., Juanchich, M., and Hagmayer, Y. (2014a). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychon. Bull. Rev*. 21, 198–204. doi: 10.3758/s13423-013-0464-6

Sirota, M., Kostovièová, L., and Juanchich, M. (2014b). The effect of iconicity of visual displays

on statistical reasoning: evidence in favor of the null hypothesis. *Psychon. Bull. Rev*. 21, 961–968. doi: 10.3758/s13423-013-0555-4

Sirota, M., Kostovičová, L., and Vallée-Tourangeau, F. (2014c). How to train your Bayesian. A problem-representation transfer rather than a format-representation shift explains training effects. *Q. J. Exp. Psychol.* doi: 10.1080/17470218.2014.972420. [Epub ahead of print].

Tabor, A., Vestergaard, C. H. F., and Lidegaard, Ø. (2009). Fetal loss rate after chorionic villus sampling and amniocentesis: an 11-year national registry study. *Ultrasound Obstet. Gynecol*. 34, 19–24. doi: 10.1002/uog.6377

Tsai, J., Miller, S., and Kirlik, A. (2011). Interactive visualizations to improve Bayesian reasoning. *Proc. Hum. Factors Ergon. Soc. Annu. Meet*. 55, 385–389. doi: 10.1177/1071181311551079

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 September 2014; accepted: 20 October 2014; published online: 06 November 2014.

Citation: Navarrete G, Correia R and Froimovitch D (2014) Communicating risk in prenatal screening: the consequences of Bayesian misapprehension. Front. Psychol. ***5****:1272. doi: 10.3389/fpsyg.2014.01272*

This article was submitted to Cognition, a section of the journal Frontiers in Psychology.

frontiers in Psychology

PERSPECTIVE
published: 17 September 2015
doi: 10.3389/fpsyg.2015.01327

Doctor, what does my positive test mean? From Bayesian textbook tasks to personalized risk communication

Gorka Navarrete[1]*, *Rut Correia*[2], *Miroslav Sirota*[3], *Marie Juanchich*[4] and *David Huepe*[1]

[1] Psychology Department, Laboratory of Cognitive and Social Neuroscience, UDP-INECO Foundation Core on Neuroscience, Universidad Diego Portales, Santiago, Chile, [2] Faculty of Education, Universidad Diego Portales, Santiago, Chile, [3] Department of Psychology, Kingston University, Kingston upon Thames, UK, [4] Department of Management, Kingston University, Kingston upon Thames, UK

Most of the research on Bayesian reasoning aims to answer theoretical questions about the extent to which people are able to update their beliefs according to Bayes' Theorem, about the evolutionary nature of Bayesian inference, or about the role of cognitive abilities in Bayesian inference. Few studies aim to answer practical, mainly health-related questions, such as, "What does it mean to have a positive test in a context of cancer screening?" or "What is the best way to communicate a medical test result so a patient will understand it?". This type of research aims to translate empirical findings into effective ways of providing risk information. In addition, the applied research often adopts the paradigms and methods of the theoretically-motivated research. But sometimes it works the other way around, and the theoretical research borrows the importance of the practical question in the medical context. The study of Bayesian reasoning is relevant to risk communication in that, to be as useful as possible, applied research should employ specifically tailored methods and contexts specific to the recipients of the risk information. In this paper, we concentrate on the communication of the result of medical tests and outline the epidemiological and test parameters that affect the predictive power of a test—whether it is correct or not. Building on this, we draw up recommendations for better practice to convey the results of medical tests that could inform health policy makers (What are the drawbacks of mass screenings?), be used by health practitioners and, in turn, help patients to make better and more informed decisions.

Keywords: Bayesian reasoning, positive predictive value, risk communication, Bayesian textbook tasks, medical tests

OPEN ACCESS

Edited by:
Bernhard Hommel,
Leiden University, Netherlands

Reviewed by:
Mark Nieuwenstein,
University of Groningen, Netherlands
Eric Johnson,
University of Barcelona, Spain

***Correspondence:**
Gorka Navarrete,
Laboratory of Cognitive and Social Neuroscience, Facultad de Psicologia, Universidad Diego Portales, Vergara 275, Santiago 8370076, Chile
gorkang@gmail.com

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *20 May 2015*
Accepted: *18 August 2015*
Published: *17 September 2015*

Citation:
Navarrete G, Correia R, Sirota M, Juanchich M and Huepe D (2015) Doctor, what does my positive test mean? From Bayesian textbook tasks to personalized risk communication. Front. Psychol. 6:1327. doi: 10.3389/fpsyg.2015.01327

Introduction

Research in Bayesian reasoning started with the pioneering work of Casscells (1978) and Eddy (1982) and has consisted mostly in asking participants about the trustworthiness of positive results in screening tests, i.e., the positive predictive value (PPV) of medical tests. The PPV of a test expresses the proportion of people affected by a medical condition relative to the total number of positive test results. Textbook Bayesian problems (as well as medical tests' brochures, informed consent forms, etc.) commonly present information about the prevalence of a condition (i.e., proportion

of population with the condition), the sensitivity of a test (i.e., probability that a test detects the presence of the medical condition) and its false-positive rate (i.e., probability that the test detects a medical condition that is not present), and ask participants to assess the positive predictive value of the test (PPV). The following example (Gigerenzer and Hoffrage, 1995) is a widely used Bayesian reasoning problem:

> *The probability of breast cancer is 1% for women aged forty who participate in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammogram. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammogram. A woman in this age group has a positive mammogram in a routine screening. What is the probability that she actually has breast cancer?*

To answer the question of PPV correctly—the probability of having the medical condition given a positive test result, formalized as p(H|D)—participants need to understand the structure of the problem and extract the key probabilistic pieces of information outlined above: the prevalence of the condition [p(H) = 1%], and the test characteristics—sensitivity (p(D|H) = 80%) and false-positive rate (p(D|∼H)=9.6%).

In this example, to adequately answer the question (PPV), a participant (or a patient) would need to combine all the above information in a specific way, following the Bayes' formula as displayed in Equation (1).

$$p\,(H|D) = \frac{p\,(H)\,p\,(D|H)}{p\,(H)\,p\,(D|H) \;+\; p\,(\sim H)\,p\,(D|\sim H)} \qquad (1)$$

Bayesian problems vary in complexity depending on the format of presentation of the probabilistic information (e.g., natural frequencies vs. single-event probability) and based on the structure and content of the narrative (Barbey and Sloman, 2007; Krynski and Tenenbaum, 2007; Lesage et al., 2013; McNair and Feeney, 2014). There are ways to simplify the computational demands: using absolute reference class (e.g., frequencies or chances with a natural sampling) and specifying the number of positive tests p(D). In this case, with only two pieces of information, p(D&H)—the chances of having a positive result and the disease at the same time—and p(D)–the chances of a positive test—we can proceed using a simplified version of the Bayes' theorem outlined in Equation (1). Equation (2) could be seen as a simple case of Laplacian probability (Laplace, 1810): *ratio of "favored events" to total possible events* (i.e., ratio of the number of correct classifications to the total positive results in the test).

$$p\,(H|D) = \frac{p\,(D\,\&\,H)}{p\,(D)} \qquad (2)$$

Researchers have found that the ability of people to solve Bayesian problems depends greatly on the way the information is conveyed, ranging from ∼5% in the first case (1), to up to ∼50% in the latter (2) (see Gigerenzer and Hoffrage, 1995 for a very detailed explanation encompassing the difference between Equations 1 and 2). Manipulating features of the textbook Bayesian problems such as visual representations (Brase, 2009; Sirota et al., 2014b), clarification of the causal structure (Krynski and Tenenbaum, 2007; McNair and Feeney, 2014), and information structure (Barbey and Sloman, 2007) can also improve reasoning performance in some circumstances. Individual differences also account for some performance variance over and above the actual content of the task, such as, for example, cognitive reflection ability and numeracy (Sirota and Juanchich, 2011; Johnson and Tubau, 2013, 2015; Lesage et al., 2013; Sirota et al., 2014a).

Furthermore, the way we currently study Bayesian reasoning may not be the best. It has been argued that research focused on how people update their beliefs or probabilities, to improve our knowledge about how the mind works, assesses ability more akin to statistical inference than to Bayesian reasoning (Mandel, 2014). But, more specifically, if we are interested in the best way to convey medical information to patients, we need to adopt a more flexible approach than the mechanical application of textbook problems. Indeed, most of the research outlined above used textbook problems to study the theoretical basis of Bayesian reasoning (Baratgin and Politzer, 2006), often using the presence of this type of information in medical contexts as a testimony of the importance of the research. The focus has been on ways to improve people's understanding via the use of pictorial aids, causal structure, computational simplification, clarification of the structure of the problem and boundary conditions (e.g., individual differences in cognitive processing), sometimes forgetting the real needs of the applied side of our research.

The importance of finding better ways to communicate medical risks has become a common motivating factor for a fair share of the Bayesian reasoning literature, given the real world impact of this field and the fact that only a few people can actually understand this kind of information as it is commonly presented (see Sedlmeier and Gigerenzer, 2001; Juslin et al., 2011; Pighin et al., 2015a). Even health-care professionals often have difficulties understanding probabilistic information[1] (Ghosh et al., 2004; Gigerenzer et al., 2007). Bayesian reasoning research has shown that people's understanding of probabilistic problems depends on the complexity of the structure of the problem, the computation required and their own cognitive skills and thinking styles. However, those principles rarely transcend the basic research walls. In clinical practice, what we know about Bayesian reasoning is not generally applied to improve the way of communicating risk. As a consequence, people have to understand their health practitioners' explanations, "informed consent" or medical tests brochures, where the information given is poorly structured, incomplete and simply often beyond their capabilities. The example below[2] shows a prenatal test brochure for Down Syndrome. As far as we have seen, this is fairly representative of the prenatal tests' brochures available online. The explanation provided in the brochure is a mix of frequencies and relative probabilities from which it is very difficult to derive the positive predictive value of the test.

[1] For example, Gigerenzer et al. (2007) show that the number of physicians able to solve a multiple choice breast cancer screening problem was 21%, slightly below chance.

[2] From http://www.prenatest.ca/en/Harmony-Prenatal-Test-Brochure.pdf.

It is estimated that trisomy 21 is present in 1 out of every 800 births in Canada.

It is estimated that trisomy 18 is present in approximately 1 out of every 6,000 births.

It is estimated that trisomy 13 is present in approximately 1 out of every 16,000 newborns.

The Harmony Test has been shown to have detection rates of up to 99 % and false positive rates as low as 0.1 % for trisomy 21, 18, and 13 (...)

In this example, if a couple expecting a baby wanted to understand what a positive result in the test meant, they would have to deal with a very complex calculation. The information given can be matched to Equation (1)—assuming you know that, p(∼H) = 1-p(H). For the trisomy 21 case it would translate into Equation (3):

$$p\left(Trisomy\ 21 \mid +test\right) =$$

$$(A)\ \frac{1\ out\ of\ 800\ \times\ 99\%}{\left(1\ out\ of\ 800{\times}99\%\right) + \left(799\ out\ of\ 800{\times}0.1\%\right)} =$$

$$(B)\ \frac{0.123}{0.123 + 0.0998} = 0.55 \qquad (3)$$

If the parents completed Equation (3)[3] they would realize the probability of having a child affected with a trisomy 21, 18, or 13 given a positive test result, is, respectively, 55%, 14% and 6% (see Navarrete et al., 2014 for a more detailed account), likely to be below their expectations, given a generally shared high regard for medical tests (Gigerenzer et al., 2009).

In a medical context, it is important that people understand the risks, the pros and cons of undertaking a test and how to interpret the result afterwards. The role of the medical personnel is vital and, although the ethical dimension and other issues involved are beyond the scope of this article, we want to recognize their complexity. In any case, we could probably agree that it is important that patients are given the possibility of reaching a sufficient level of understanding to give a truly informed consent. Why then are we forcing participants and patients to deal with a non-trivial set of information, and then to perform a calculation generally too difficult for them? In most cases this translates into patients or doctors being unable to provide an informed consent and to blindly trusting medical tests or falling prey to bogus medical tests, and in uninformed politicians implementing policies promoting mass screenings for low prevalence diseases, where the positive predictive value is also low (e.g., as for the Trisomy 13 for which a positive test identifies correctly the Syndrome in only 6 cases out of 100). This can result in negative consequences, costing life and money (Gigerenzer et al., 2007).

But why are mass screenings less useful than targeted screenings? To be able to understand the result of a medical test, one needs to take into account two different and inter-related sets of information. The first set of information relies on the test's characteristics: its sensitivity and false positive rate. The second set of information has to do with the disease itself, more specifically its prevalence. The usefulness and trustworthiness of a test critically depends on the prevalence of the medical condition it is seeking to detect, and this depends on the reference group used (Baldessarini et al., 1983).

Prevalence,– and its relationship with false positives– is pivotal and very often misunderstood when interpreting the meaning of a positive result in a test. As prevalence decreases—as is the case in mass screenings—even near perfect tests produce a large number of false positives, and hence, a low PPV. Several authors have warned about the dangers of mass screenings and their negative consequences, such as the high cost of false positives in psychological and monetary terms (Christiansen et al., 2000; Gigerenzer et al., 2007; Navarrete et al., 2014).

It is important to keep in mind that prevalence is not a characteristic of a test but of the population to whom the test is given. For example, the prevalence of certain chromosomal aberrations in fetuses is related to maternal age and gestation time (Nicolaides, 2004). The exact same test would "work" a lot better—i.e., have a higher PPV—in older pregnant women than in younger ones. Specifically, the rates of prevalence range from 1 out of 1000 for 20 year old mothers up to 1 in 38 for 42 year old mothers (Nicolaides, 2004, p. 18). That means that the combined test reliability, used commonly as a screening procedure, goes from a 2% PPV when used in young mothers to 34% PPV when used in a relatively high risk group. Still a far cry from a reliable test, but a change with dramatic consequences given the default recommended assessment in the case of a positive result, and its associated risks (Navarrete et al., 2014).

To combine all available information, one should follow Equation (2): ratio of the number of correct classifications to the total positive results in the test. The number of correct classifications will always be close to 1 as the prevalence is usually presented in a standard way—*1* out of X (but see Pighin et al., 2015b for some related issues)—and the sensitivity is usually close enough to 100%. On the other hand, the denominator magnitude will depend on the number of false positives and the *X* term of the prevalence (1 out of *X*). Imagine we have a test with a 0.1% rate of false positives that aims to detect a relatively common condition affecting 1 in 100 individuals (see Equation 4). The number of false positives would be calculated multiplying the 99 healthy individuals by 0.1%, that is, 99 × 0.001=∼ 0.099. Using Equation (2), this would translate into a PPV of 0.91, or a 91% chance of having the medical condition given a positive test result.

$$p\left(H|D\right) = \frac{p\left(D\&H\right)}{p\left(D\right)} = \frac{1}{1 + 0.099} = 0.91 \qquad (4)$$

Unfortunately, tests are not always so reliable, nor are the tested medical conditions so common. According to the EU regulations[4], most patients suffer from diseases affecting 1 in 100,000. Test reliability and prevalence can dramatically reduce

[3]To be able to give a reference point, we asked 66 people to solve the above two Equations 3(A) and 3(B) through the web platform Amazon's Mechanical Turk, and the average accuracy correct response was 21 and 53%, respectively. That is, even when we give people the data of the brochure within the required formula, less than 25% are able to correctly solve it.

[4]From http://ec.europa.eu/health/rare_diseases/policy/index_en.htm: "In EU countries, any disease affecting fewer than five people in 10,000 is considered rare.

the ability of a test to identify a medical condition. For example, a test with the same rate of false positives (0.1%) that aims to detect a disease with a lower incidence, such as of 1 in 10,000 would result in a much lower PPV: 0.09, or 9%, as seen in Equation (5).

$$p\,(H|D) = \frac{p\,(D\&H)}{p\,(D)} = \frac{1}{1+9.99} = 0.09 \tag{5}$$

The previous two examples show how the PPV of a test can change from 91 to 9% simply because of a lower incidence of a medical condition (from 1 in 100 to 1 in 10,000). In a mass screening campaign, the incidence of a medical condition is lower than in a targeted screening campaign, lowering dramatically the reliability of the test results.

Of course, as often happens, if a medical test is not as reliable as the one used in the two examples above (100% sensitivity, and 0.1% false positive rate), a low positive predictive value appears even with common medical conditions. For example, see in Equation (6) the computation of the positive predictive value of a test aiming to detect a condition with a prevalence of 1 in 100, and a false positive rate as low as 1%. When the rate of false positives increases by 0.9%, the positive predictive value of the test decreases by 40%, dropping from 90 to 50%. In this context, a person receiving a positive test has only a 50% chance of actually having the condition.

$$p\,(H|D) = \frac{p\,(D\&H)}{p\,(D)} = \frac{1}{1+0.99} = 0.5 \tag{6}$$

That number may seem small, but it translates into approximately 246,000 people throughout the EU's 28 member countries. Most patients suffer from even rarer diseases affecting one person in 100,000 or more. It is estimated that today in the EU, 5-8,000 distinct rare diseases affect 6–8% of the population—between 27 and 36 million people." The PPV for a test with 100% sensitivity and a 0.1% false positive rate trying to detect a 1 in 2000 condition is 50%.

With all these examples, we are not implying that screening tests should not be trusted. We intend to outline the factors needed to be considered when using and interpreting medical test results. As we have seen, low prevalence rates, and their interaction with false positive rates, are generally guilty of decreasing the positive predictive value of a test: **Figure 1** provides an illustration of this. The variability of positive predictive values of medical tests, according to the characteristics of the test and the prevalence of the condition, makes it hard for patients to decide whether to take the test and to assess their chances of having a condition when they test positive, particularly when the information given to them is generally too complicated to understand.

Given the need of facilitating the patient's assessment and decision making powers, different solutions can be offered. Further medical research to improve the present tests and decrease their false positive rates is obviously a very important and necessary path. Testing only people in higher risk groups and avoiding mass screenings as much as possible or, at least, making their limitations clear, is a critical necessity given the reality of the medical tests available and their trustworthiness for diagnosing rare conditions. Of course, increasing public health literacy should be traversal to these and any other alternatives available (Gigerenzer, 2015).

Nonetheless, one important aspect not covered in the above options is that we need to find better ways to communicate medical risks, starting with using the information obtained through empirical research in medical practice. For those of us interested in improving the way we convey medical risks, focusing research on what real patients need is vital. In the real world, when receiving medical test results or reading informed consents, people are confronted with probabilistic information generally too complex to be understood, let alone calculated. We need to avoid altogether the classical triad (specificity, false

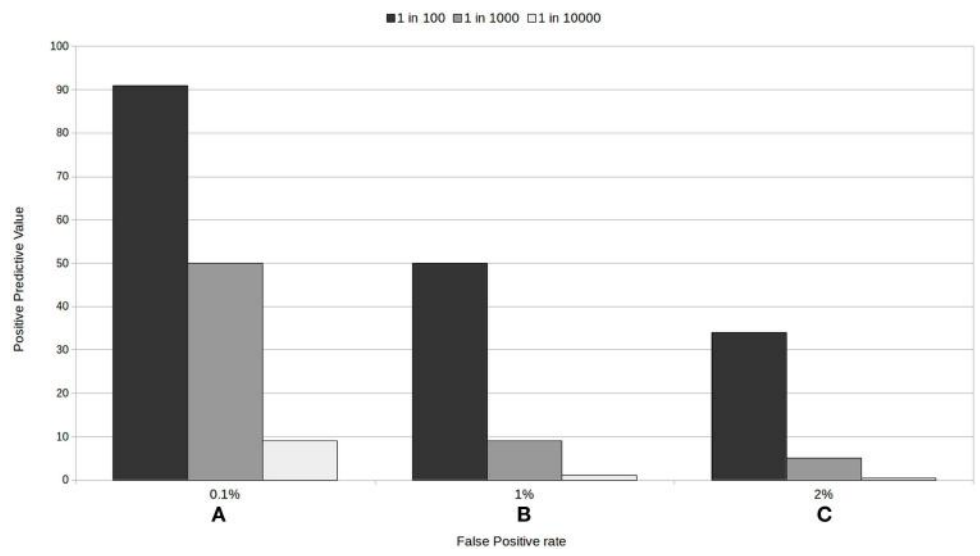

FIGURE 1 | Positive predictive value for three tests with a 100% sensitivity according to the rate of false positive (A) 0.1%, (B) 1%, and (C) 2%, and to the prevalence of the condition.

positive rate and prevalence) if we want to improve people's chances of understanding test results and informed consents, and of playing a more active role in shared decision making. It is also important to acknowledge that there exist teams focusing on helping health practitioners better communicate risk and patients better understand risks (e.g., Reyna et al., 2009; Garcia-Retamero et al., 2010; Gigerenzer, 2014). However, theoretical research seems to still have the lion's share in Bayesian reasoning and we would suggest further harnessing these teams' work to derive simple and effective guidelines to communicate medical test results.

Our proposal, then, is to present information about the PPV, and specifically, how trustworthy a positive or a negative result in each particular test really is *for the individual*: that is, the PPV for the test relative to the risk group the person belongs to. Using epidemiological factors (such as age in the prenatal screening example above, a list of common behaviors for each risk group, family history, etc.) we could help people assign themselves to a specific risk group. An example would be to present something akin to one of the sections of **Figures 1A–C**, making clear which epidemiological factors, risk behaviors, etc. are associated with each of the prevalence or risk groups. In prenatal screening, this would depend, amongst other factors, on the age of the mother to be. In a mass screening context, this approach could translate to most people (low risk people) avoiding getting tested for rare conditions, as the PPV for them would be extremely low. Prevalence is a characteristic of the disease or of the group tested and its risk factors, and not of the test, and we must stop ignoring this fact. This would help people distinguish between good and bad tests and make for more informed decisions.

To sum up, the goal of this article is to call on the scientific community studying Bayesian reasoning to join efforts and focus further on finding better ways to present medical information. Such research could inform policy makers' decisions (specifically helping them understand why mass screenings are less useful than targeted screenings) and be used by health staff to enable patients to make better informed decisions related to their health. One possibility is to find good ways to assign people to risk groups and to present information about tests relative to these risk groups, but other options surely exist. Of course, it is important to empirically confirm that people really do better with this new way of presenting the information (e.g., they do understand the pros and cons of the combination of tests suggested in prenatal screening), and to assess the medical consequences of such trials. This call for further applied research is not unique and joins other initiatives to avoid risk miscommunication (e.g., fact-box for breast cancer screening pamphlets as suggested by Gigerenzer, 2014). Most people would agree: *misinformation needs to stop*. We have the chance to work toward this goal together.

Acknowledgments

This work was supported by grants from Comisión Nacional de Investigación Científica y Tecnológica (CONICYT/FONDECYT Regular 1150824 to GN and 1140114 to DH); and the Semilla grants from the University Diego Portales (SEMILLA201418 to GN). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We also want to thank Sarah Nuttall for her help proofreading the manuscript.

References

Baldessarini, R. J., Finklestein, S., and Arana, G. W. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. *Arch. Gen. Psychiatry*, 40, 569–573. doi: 10.1001/archpsyc.1983.01790050095011

Baratgin, J., and Politzer, G. (2006). Is the mind Bayesian? The case for agnosticism. *Mind Society* 5, 1–38. doi: 10.1007/s11299-006-0007-1

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

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol.* 23, 369–381. doi: 10.1002/acp.1460

Casscells, W. (1978). Interpretation by physicians of clinical laboratory results. *N.Engl. J. Med.* 299, 999–1001. doi: 10.1056/NEJM197811022991808

Christiansen, C. L., Wang, F., Barton, M. B., Kreuter, W., Elmore, J. G., Gelfand, a, E., and Fletcher, S. W. (2000). Predicting the cumulative risk of false-positive mammograms. *J. Natl. Cancer Inst.* 92, 1657–1666. doi: 10.1093/jnci/92.20.1657

Eddy, D. (1982). "Probabilistic reasoning in clinical medicine: problems and opportunities," in *Judgment under Uncertainty: Heuristics and Biases,* eds D. Kahneman, P. Slovic and A. Tversky (Cambridge: Cambridge University Press), 249–267.

Garcia-Retamero, R., Galesic, M., and Gigerenzer, G. (2010). Do icon arrays help reduce denominator neglect? *Med. Decis. Making* 30, 672–684. doi: 10.1177/0272989X10369000

Ghosh, A. K., Ghosh, K. and Erwin, P. J. (2004). Do medical students and physicians understand probability? *QJM* 97, 53–55. doi: 10.1093/qjmed/hch010

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychol. Sci. Public Interest* 8, 53–96. doi: 10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., Mata, J., and Frank, R. (2009). Public knowledge of benefits of breast and prostate cancer screening in Europe. *J. Natl. Cancer Inst.* 101, 1216–1220. doi: 10.1093/jnci/djp237

Gigerenzer, G. (2014). Breast cancer screening pamphlets mislead women. *BMJ* 2636, 1–2. doi: 10.1136/bmj.g2636

Gigerenzer, G. (2015). Towards a paradigm shift in cancer screening: informed citizens instead of greater participation. *BMJ* 350, 1–2. doi: 10.1136/bmj.h2175

Johnson, E. D., and Tubau, E. (2013). Words, numbers, & numeracy: diminishing individual differences in Bayesian reasoning. *Learn. Individ. Differ.* 28, 34–40. doi: 10.1016/j.lindif.2013.09.004

Johnson, E. D., and Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Front. Psychol.* 6:938. doi: 10.3389/fpsyg.2015.00938

Juslin, P., Nilsson, H., Winman, A., and Lindskog, M. (2011). Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats. *Cognition* 120, 248–267. doi: 10.1016/j.cognition.2011.05.004

Krynski, T. R., and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *J. Exp. Psychol. Gen.* 136, 430–450. doi: 10.1037/0096-3445.136.3.430

Laplace, P. (1810). *Analytic Theory of Probabilities*. Paris: Imprimerie Royale.

Lesage, E., Navarrete, G., and De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: the role of general cognitive resources. *Think. Reason.* 19, 27–53. doi: 10.1080/13546783.2012.713177

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Front. Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

McNair, S., and Feeney, A. (2014). When does information about causal structure improve statistical reasoning? *Q. J. Exp. Psychol.* 67, 625–645. doi: 10.1080/17470218.2013.821709

Navarrete, G., Correia, R., and Froimovich, D. (2014). Communicating risk in prenatal screening: the consequences of Bayesian misapprehension. *Front. Psychol.* 5:1272. doi: 10.3389/fpsyg.2014.01272

Nicolaides, K. H. (2004). *The 11–13 +6 Weeks Scan*. London: Fetal Medicine Foundation.

Pighin, S., Gonzalez, M., Savadori, L., and Girotto, V. (2015a). Improving public interpretation of probabilistic test results: distributive evaluations. *Med. Decis. Making* 35, 12–15. doi: 10.1177/0272989X14536268

Pighin, S., Savadori, L., Barilli, E., Galbiati, S., Smid, M., Ferrari, M., et al. (2015b). Communicating Down syndrome risk according to maternal age: "1-in-X" effect on perceived risk. *Prenat. Diagn.* 35, 777–782. doi: 10.1002/pd.4606

Reyna, V. F., Nelson, W. L., Han, P. K., and Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychol. Bull.* 135, 943–973. doi: 10.1037/a0017327

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Sirota, M., and Juanchich, M. (2011). Role of numeracy and cognitive reflection in Bayesian reasoning with natural frequencies. *Stud. Psychol.* 53, 151–161.

Sirota, M., Juanchich, M., and Hagmayer, Y. (2014a). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychon. Bull. Rev.* 21, 198–204. doi: 10.3758/s13423-013-0464-6

Sirota, M., Kostovicová, L., and Juanchich, M. (2014b). The effect of iconicity of visual displays on statistical reasoning: evidence in favor of the null hypothesis. *Psychon. Bull. Rev.* 21, 961–968. doi: 10.3758/s13423-013-0555-4

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in
PSYCHOLOGY

OPINION ARTICLE
published: 09 October 2014
doi: 10.3389/fpsyg.2014.01144

The psychology of Bayesian reasoning

David R. Mandel*

Socio-Cognitive Systems Section, Defence Research and Development Canada and Department of Psychology, York University, Toronto, ON, Canada
**Correspondence: david.mandel@drdc-rddc.gc.ca*

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
Vittorio Girotto, University IUAV of Venice, Italy
Miroslav Sirota, King's College London, UK

Keywords: Bayesian reasoning, belief revision, subjective probability, human judgment, psychological methods

Most psychological research on Bayesian reasoning since the 1970s has used a type of problem that tests a certain kind of statistical reasoning performance. The subject is given statistical facts within a hypothetical scenario. Those facts include a base-rate statistic and one or two diagnostic probabilities. The subject is meant to use that information to arrive at a "posterior" probability estimate. For instance, in one well-known problem (Eddy, 1982) the subject encounters the following:

> The probability of breast cancer is 1% for a woman at age forty who participates in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? __ %.

The information in such problems can be mapped onto common expressions that use H as the focal hypothesis, $\neg H$ as the mutually-exclusive hypothesis, and D as datum: $P(H)$, the prior (often equated with the base-rate) probability of the hypothesis; $P(D|H)$, the true-positive rate; and $P(D|\neg H)$, the false-positive rate. In the mammography problem, $P(H) = 0.01$, $P(D|H) = 0.80$, and $P(D|\neg H) = 0.096$. Furthermore, $P(\neg H) = 1 - P(H) = 0.99$. The estimate queried is $P(H|D)$.

Bayes' theorem states:

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

Thus, it yields a posterior probability of 0.078 in the mammography problem. Yet even the majority of physicians who were queried by Eddy (1982) gave estimates roughly one order of magnitude higher (i.e., 0.70–0.80).

Well-established findings such as these have supported the view that expert and naïve subjects alike are non-Bayesian (Kahneman and Tversky, 1972). A common explanation is that people neglect base-rate information, which is not tracked by the intuitive heuristics they use to reach an estimate (Kahneman and Tversky, 1972, 1973). For instance, if base rates were neglected in the mammography problem,

$$P(H|D) = \frac{0.80}{0.80 + 0.096} \approx 0.89.$$

This estimate is closer to the modal estimate but is still off by about ten percentage points. Another explanation is that people commit the *inverse fallacy*, confusing $P(H|D)$, which they are asked to estimate, with $P(D|H)$, which is provided (Koehler, 1996). In the mammography problem, this explanation fits the data well because $P(D|H) = 0.80$. The inverse fallacy can also explain patterns of deviation from Bayes' theorem in tasks that hold constant base rates for alternative hypotheses (Villejoubert and Mandel, 2002).

It is also known that steps can be taken to increase agreement with Bayes' theorem. Since Bayes' theorem can be simplified as

$$P(H|D) = \frac{f(D \cap H)}{f(D)},$$

task reformulations that directly provide these values or make them easily computable increase the proportion of Bayesian responses (e.g., Gigerenzer and Hoffrage, 1995; Hoffrage et al., 2002; Ayal and Beyth-Marom, 2014). Such formulations of evidence reduce computational steps and may also effectively trigger awareness of the correct solution, much as eliciting logically-related probability estimates (e.g., of binary complements) in close proximity rather than far apart improves adherence to the additivity property (Mandel, 2005; Karvetski et al., 2013). Natural frequency representations, which reveal nested-set relations among a reference class or representative sample (Gigerenzer and Hoffrage, 1995; Cosmides and Tooby, 1996), lend themselves easily to such simplification and have been shown to improve Bayesian reasoning. For instance, Bayesian responses to the mammography problem more than doubled when it was presented in natural-frequency format (Gigerenzer and Hoffrage, 1995). Although the theoretical bases of such improvements are debated (e.g., Barbey and Sloman, 2007, and continuing commentaries), most agree that substantial improvement in conformity to Bayes' theorem is achievable in this manner.

Bayesian reasoning also benefits from the use of visual representations of pertinent statistical information, such as Euler circles (Sloman et al., 2003) and frequency grids or trees (Sedlmeier and Gigerenzer, 2001), which further clarify nested-set relations. For instance, **Figure 1** shows how the natural-frequency version of the mammography problem could be represented with a frequency tree to help

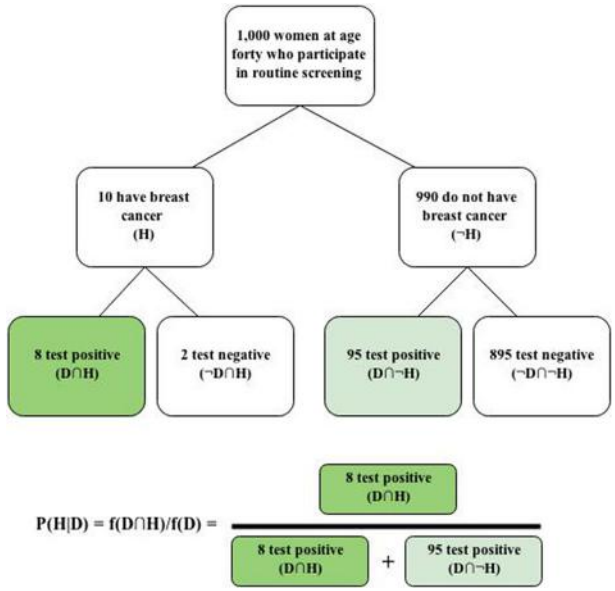

FIGURE 1 | Frequency tree and solution for the mammography problem.

individuals visualize the nested-set relations and how such information ought to be used to compute the posterior probability.

OBSERVATIONS

A remarkable feature of the standard approach to studying Bayesian reasoning is its inability to reveal how people *revise* their beliefs or subjective probabilities in light of newly acquired evidence. That is, in tasks such as the mammography problem, information acquisition is not staged across time (real or hypothetical), and researchers typically do not collect multiple "prior" and "posterior" (i.e., revised) probability assessments.

It is instead conveniently assumed that the base rate represents the subject's prior belief, $P(H)$, which the subject updates in light of "new" evidence, D. It is somewhat ironic that advocates of base-rate neglect have not noted (let alone warned) that, if people ignore base rates, it may be unwise to assume they represent the subject's prior probability. Would that not imply that the subject ignores his or her own prior probability?

Priors need not equal base rates, as many have noted (e.g., de Finetti, 1964; Niiniluoto, 1981; Levi, 1983; Cosmides and Tooby, 1996). The prior, $P(H)$, is in fact a *conditional* probability corresponding to one's personal probability of H, given all that they know prior to learning D (Edwards et al., 1963; de Finetti, 1972). In all real-life cases where no single, relevant base rate is ever explicitly provided, people may experience considerable uncertainty and difficulty in deciding precisely which base rate is the most relevant one to consider. For instance, imagine that the test result in the mammography problem is for a specific, real woman and not just an abstract one lacking in other characteristics. If her prior for H is contingent on the presence or absence of some of those characteristics, one could see how the base rate provided in the problem might be more or less relevant to the woman's particular case. If she has several characteristics known to elevate a woman's risk of breast cancer, then simply using the base rate for 40-year-old women as her prior would bias her revised assessment by leading her to underestimate the risk she faces. Conversely, she may have a configuration of characteristics that make her less likely than the average 40-year-old woman to develop breast cancer, in which case using the base rate as her prior would cause her to overestimate objective risk.

Clearly, the ideal base rate in such personal cases would be a sample of people who are just like the patient, yet since each of us is unique no such sample exists. In the absence of a single, ideal base rate, one must decide among a range of imperfect ones—a task involving decision under uncertainty. It might be sensible for the woman getting the screening to anchor on a relevant, available base rate, such as for women in her cohort, and then adjust it in light of other diagnostic characteristics that she knows she possesses. Yet, if people are overly optimistic (Taylor and Brown, 1988; Weinstein, 1989), we might anticipate systematic biases in adjustment, with underweighting of predisposing factors and overweighting of mitigating factors. This point about the possible role of motivated cognition also brings a key tenet of subjective Bayesianism to the fore—namely, that different individuals with access to the same information could have different degrees of belief in a given hypothesis, and they may be equally good Bayesians as long as they are equally respectful of static and dynamic coherence requirements (Baratgin and Politzer, 2006).

Given that standard Bayesian reasoning tasks involve no assessment of a prior probability, they should be seen for what they are: conditional probability judgment tasks that require the combination of statistical information. When that information is fleshed out, it reveals the fours cells of a 2×2 contingency table, where $a = f(H \cap D)$, $b = f(H \cap \neg D)$, $c = f(\neg H \cap D)$, and $d = f(\neg H \cap \neg D)$. Going from left to right, the four boxes in the lowest level of the frequency tree in **Figure 1** correspond to cells *a–d,* which have received much attention in the causal induction literature (Mandel and Lehman, 1998). We can restate Bayes' theorem as the following cell-frequency equalities, corresponding to short and long expressions given earlier, respectively:

$$P(H|D) = \frac{a}{a+c}$$

$$= \frac{\begin{array}{c}(a+b)/(a+b+c+d) \times \\ a/(a+b)\end{array}}{\begin{array}{c}(a+b)/(a+b+c+d) \times \\ a/(a+b) + (c+d)/(a+b+c+d) \\ \times c/(c+d)\end{array}}.$$

From this perspective, it is perhaps unsurprising why a greater proportion of subjects conform to Bayes theorem when they are given the frequencies *a–d* than when they are instead given the values equal to $(a+b)/(a+b+c+d)$, $a/(a+b)$, and

$c/(c+d)$. That is, frequencies *a–c* support the easy computation of $a/(a+c)$. However, those improvements in performance, which pertain to static coherence constraints (Baratgin and Politzer, 2006), do not speak to other important facets of Bayesian reasoning, such adherence to dynamic coherence constraints, which are fundamental to Bayesian belief revision (Seidenfeld, 1979).

I do not intend for my observations to imply that the well-established findings I summarized earlier are incorrect. However, I believe greater care should be taken in labeling the type of performance measured in such experiments. "Statistical inference" would seem to be more appropriate than "Bayesian reasoning" given the limitations I have noted.

Future research on Bayesian reasoning would benefit from a richer conceptualization of what it is to "be Bayesian" and from better discussion of whether being non-Bayesian is necessarily irrational (Lewis, 1976; Walliser and Zwirn, 2002; Baratgin and Politzer, 2006). Future work would also benefit by breaking free of the typical methodological approach exemplified by the mammography problem. One avenue would be to collect prior and posterior assessments from subjects in experiments where information acquisition is staged (e.g., Girotto and Gonzalez, 2008), or where temporal staging is at least an important characteristic of the described problem, such as in the Monty Hall problem (Krauss and Wang, 2003) and Sleeping Beauty problem (Elga, 2000; Lewis, 2001). Another promising line involves assessing people's prior distributions for different types of real events (e.g., Griffiths and Tennenbaum, 2006).

The staging of information with repeated assessments was in fact a common methodological approach in Bayesian research prior to the 1970s, culminating in the classic work on conservatism by Ward Edwards and others (for a review, see Slovic and Lichtenstein, 1971). Such approaches could be revisited in new forms and contrasted with other methods of information staging, such as the trial-by-trial information acquisition designs used in causal induction (e.g., Kao and Wasserman, 1993; Mandel and Vartanian, 2009) or category learning (e.g., Gluck and Bower, 1988; Shanks, 1990) studies.

For example, Williams and Mandel (2007) presented subjects with 28 problems prompting them for a conditional probability judgment. In each problem, subjects first saw 20 patient results presented serially. The subject saw whether the patient carried a virus hypothesized to cause a particular illness and whether the patient had the illness or not. Sample characteristics were varied so that $P(H|D)$ ranged from 0 to 1 over seven probability levels across the problems. Subjects exhibited a form of conservatism (cf. Edwards, 1968), overestimating low probabilities and underestimating high probabilities. The task illustrates the value of breaking free of the standard problem set. First, the trial-by-trial design better represents the information acquisition environment that ecological rationality theorists (e.g., Gigerenzer and Hoffrage, 1995; Cosmides and Tooby, 1996), have described as natural. That is, information acquisition in that task is more natural than in natural-frequency versions of standard problems because no statistical information is presented to the subject in written form. Rather, subjects learn about each case serially, more like they would have in the Paleolithic Era. Second, the design gets researchers away from studying average responses to a single problem with a unique data configuration. The authors would not have been able to detect conservatism if they had not explored problems for which the mathematical probabilities subjects were asked to judge covered the full probability range. Third, the induction paradigm, which presents information on cells *a–d* to subjects, easily lends itself to studying subjective cell importance, which can help take the cognitive processes subjects use to arrive at their judgments out of the proverbial black box. For instance, Williams and Mandel (2007) found that, when asked to assign subjective importance ratings to each of the fours cells, subjects assigned weight to irrelevant information, such as focusing on $\neg D$ cases when asked to judge $P(H|D)$, causing an underweighting of relevant information.

The issues I have raised, non-exhaustive as they are, draw attention to some important problems with the conventional approach to studying Bayesian reasoning in psychology that has been dominant since the 1970s. Rather than fostering pessimism, I hope my comments illustrate that there are good opportunities for future work to advance our understanding of how people revise or update their beliefs.

ACKNOWLEDGMENTS

I thank Baruch Fischhoff, Vittorio Girotto, Gorka Navarrete, and Miroslav Sirota for helpful comments on earlier drafts of this paper.

REFERENCES

Ayal, S., and Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgm. Decis. Mak.* 9, 226–242.

Baratgin, J., and Politzer, G. (2006). Is the mind Bayesian? The case for agnosticism. *Mind Soc.* 5, 1–38. doi: 10.1007/s11299-006-0007-1

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–297. doi: 10.1017/S0140525X07001653

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

de Finetti, B. (1964). "Foresight: its logical laws, its subjective sources," in *Studies in Subjective Probability* (1st Edn., 1937), eds H. E. Kyburg and H. E. Smokler (New York, NY: Wiley), 53–118.

de Finetti, B. (1972). "Probability, statistics and induction: their relationship according to the various points of view," in *Probability, Induction and Statistics. The Art of Guessing* (1st Edn., 1959), ed B. de Finetti (London: Wiley), 141–228.

Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine: problems and opportunities," in *Judgment under Uncertainty: Heuristics and Biases*, eds D. Kahneman, P. Slovic and A. Tversky (New York, NY: Cambridge University Press), 249–267. doi: 10.1017/CBO9780511809477.019

Edwards, W. (1968). "Conservatism in human information processing," in *Formal Representation of Human Judgment*, ed B. Kleinmuntz (New York, NY: Wiley), 17–52.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* 70, 193–242. doi: 10.1037/h0044139

Elga, A. (2000). Self-locating belief and the sleeping Beauty problem. *Analysis* 60, 143–147. doi: 10.1093/analys/60.2.143

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Girotto, V., and Gonzalez, M. (2008). Children's understanding of posterior probability. *Cognition* 106, 325–344. doi: 10.1016/j.cognition.2007.02.005

Gluck, M. A., and Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *J. Exp. Psychol. Gen.* 117, 227–247. doi: 10.1037/0096-3445.117.3.227

Griffiths, T. L., and Tennenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychol. Sci.* 17, 767–773. doi: 10.1111/j.1467-9280.2006.01780.x

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition* 84, 343–352. doi: 10.1016/S0010-0277(02)00050-1

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychol. Rev.* 80, 237–251. doi: 10.1037/h0034747

Kao, S.-F., and Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 1363–1386. doi: 10.1037/0278-7393.19.6.1363

Karvetski, C. W., Olson, K. C., Mandel, D. R., and Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decis. Anal.* 10, 305–326. doi: 10.1287/deca.2013.0279

Koehler, J. J. (1996). The base rate fallacy reconsidered: descriptive, normative and methodologica l challenges. *Behav. Brain Sci.* 19, 1–53. doi: 10.1017/S0140525X00041157

Krauss, S., and Wang, X. T. (2003). The psychology of the Monty Hall Problem: discovering psychological mechanisms for solving a tenacious brain teaser. *J. Exp. Psychol. Gen.* 132, 3–22. doi: 10.1037/0096-3445.132.1.3

Levi, I. (1983). Who commits the base rates fallacy. *Behav. Brain. Sci.* 6, 502–506. doi: 10.1017/S0140525X00017209

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philos. Rev.* LXXXV, 297–315. doi: 10.2307/2184045

Lewis, D. (2001). Sleeping beauty: reply to Elga. *Analysis* 61, 171–176. doi: 10.1093/analys/61.3.171

Mandel, D. R. (2005). Are risk assessments of a terrorist attack coherent? *J. Exp. Psychol. Appl.* 11, 277–288. doi: 10.1037/1076-898X.11.4.277

Mandel, D. R., and Lehman, D. R. (1998). Integration of contingency information in judgments of cause, covariation, and probability. *J. Exp. Psychol. Gen.* 127, 269–285. doi: 10.1037/0096-3445.127.3.269

Mandel, D. R., and Vartanian, O. (2009). Weighting of contingency information in causal judgment: evidence of hypothesis dependence and use of a positive-test strategy. *Q. J. Exp. Psychol.* 62, 2388–2408. doi: 10.1080/17470210902794148

Niiniluoto, I. (1981). Cohen versus Bayesianism. *Behav. Brain. Sci*. 4, 349. doi: 10.1017/S0140525X00009274

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Seidenfeld, T. (1979). Why I am not an objective Bayesian: some reflections prompted by Rosenkrantz. *Theory Decis.* 11, 413–440. doi: 10.1007/BF00139451

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Q. J. Exp. Psychol.* 42A, 209–237. doi: 10.1080/14640749008401219

Sloman, S. A., Over, D. E., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Process.* 91, 296–309. doi: 10.1016/S0749-5978(03)00021-9

Slovic, P., and Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organ. Behav. Hum. Perform.* 6, 649–744. doi: 10.1016/0030-5073(71)90033-X

Taylor, S. E., and Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychol. Bull.* 103, 193–210. doi: 10.1037/0033-2909.103.2.193

Villejoubert, G., and Mandel, D. R. (2002). The inverse fallacy: an account of deviations from Bayes's theorem and the additivity principle. *Mem. Cogn.* 30, 171–178. doi: 10.3758/BF03195278

Walliser, B., and Zwirn, D. (2002). Can Bayes' rule be justified by cognitive rationality principles? *Theory Decis.* 53, 95–135. doi: 10.1023/A:1021227106744

Weinstein, N. D. (1989). Optimistic biases about personal risks. *Science* 264, 1232–1233. doi: 10.1126/science.2686031

Williams, J. J., and Mandel, D. R. (2007). "Do evaluation frames improve the quality of conditional probability judgment?," in *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, eds D. S. McNamara and J. G. Trafton (Mahwah, NJ: Erlbaum), 1653–1658.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 September 2014; accepted: 19 September 2014; published online: 09 October 2014.

Citation: Mandel DR (2014) The psychology of Bayesian reasoning. Front. Psychol. ***5****:1144. doi: 10.3389/fpsyg.2014.01144*

This article was submitted to Cognition, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

OPINION ARTICLE
published: 06 February 2015
doi: 10.3389/fpsyg.2015.00097

Beyond the status-quo: research on Bayesian reasoning must develop in both theory and method

Simon J. McNair*

Centre for Decision Research, Leeds University Business School, University of Leeds, Leeds, UK
**Correspondence: s.j.mcnair@leeds.ac.uk*

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
Eric Johnson, University of Barcelona, Spain
Marie Juanchich, Kingston Business School, UK

Keywords: Bayesian reasoning, individual differences, cognition, psychological methods, subjective probability

Judgements in the real-world often inherently involve uncertainty, from the mundane: "do those clouds signal rain?" to the potentially life-changing: "Does this person have cancer?" Normatively estimating the likelihood of outcomes in such situations involves considering how competing sources of probabilistic evidence ("how likely are clouds with/without rain?") should be weighed against prior probabilities ("how likely is it to rain/not rain?"), known as *Bayesian reasoning*. This complex form of reasoning, however, typically eludes many people, and can have dramatic implications including overdiagnosis (e.g., Casscells et al., 1978), and wrongful conviction (e.g., the famous Sally Clark case in the UK. See Nobles and Schiff, 2007). Whilst the question of how best to assist people to make such judgments remains in critical need of research (e.g., Navarrete et al., 2014), this paper considers how extant research on Bayesian facilitation has been somewhat constrained by both theoretical, and methodological status-quos. As Mandel (2014) notes, in more general terms we still know relatively little about "what it is to 'be Bayesian'," which has clear implications for our understanding of "what works and why" in Bayesian intervention. This paper contemplates several suggestions as to how research may improve its pursuit of this goal, including the deconstructing of Bayesian reasoning into component tasks, and the leveraging of more process-oriented measures to further integrate burgeoning findings concerning individual cognitive differences.

Although research has discovered several interventions that can facilitate more accurate Bayesian judgments, discussion has centered on a distinct division as to the psychological basis of these facilitation effects. Facilitation is often explained as being due to either (a) humans having evolved a cognitive primacy specifically for naturally sampled data (e.g., Gigerenzer and Hoffrage, 1995; Brase, 2009), or alternatively (b) an activation of more general analytical cognitive processes through explicating nested subset relations (e.g., Sloman et al., 2003; Yamagishi, 2003). Whilst the former, evolutionary hypothesis advocates facilitation through expressing data as natural frequencies, the latter, nested-sets hypothesis argues that reasoning can be improved irrespective of numerical format by generally clarifying set relations in the structure of the available evidence, such as through the use of visual diagrams. The debate between both positions, to a large extent, continues to define the literature on Bayesian reasoning (more recently Brase, 2008; Hill and Brase, 2012; Lesage et al., 2013; Sirota et al., 2014). But, whilst there continues to be disagreement on how best to facilitate Bayesian reasoning, one might look to the research and note the distinct variability in reported improvements produced by both frequency- and set-based interventions.

To illustrate, uncertain data expressed as naturally-sampled frequencies can increase Bayesian accuracy as high as either 76% (Cosmides and Tooby, 1996), 54% (Evans et al., 2000), or 31% (Sloman et al., 2003) where equivalent measures have been used. Similarly, equivalent visual diagrams that elucidate nested set relations, irrespective of numerical format, can improve accuracy rates as high as 80% (Yamagishi, 2003), 48% (Sloman et al., 2003), or 35% (Brase, 2009). Such variability exposes a particular limitation common to both perspectives in that neither theory offers satisfactory explanations as to why many people are seemingly *not* facilitated by their respective interventions. This perhaps stems more generally from the fact that both perspectives provide little specification of the actual mental journey people undergo when attempting to reason in Bayesian terms. By more clearly characterizing what distinguishes those who are and those who are not facilitated we might overcome some of these theoretical limitations and, ultimately, further extend our understanding of how best to improve Bayesian reasoning beyond the theoretical divide that currently exists.

Approaching this issue involves a slight shift in perspective from "what works and why?" in Bayesian facilitation to "what works for whom, and why?" (see Hill and Brase, 2012; McNair and Feeney, in press, for examples), and more recent research has begun to illuminate a diverse range of psychological capacities associated with Bayesian facilitation. Abilities such as numeracy (e.g., Johnson and Tubau, 2013; McNair and Feeney, in press; though see also Hill and Brase, 2012); cognitive reflection (Lesage et al., 2013); and fluid intelligence (e.g., Sirota et al., 2014) have variously being associated with good Bayesian reasoning, which may go some way in explaining why previous research has noted such variability in facilitation findings (see Brase et al., 2006, for related concerns). Yet, identifying that component

abilities and traits are associated with facilitation effects answers only part of the above question. Moreover, recent discussion of individual differences in Bayesian facilitation has remained grounded in the evolutionary and nested-sets debate as it stands, and as such there exists limited extrapolation of these findings beyond the abstract activation of either a frequency-processing engine in the brain, or set-based analytical processing [though see discussions of Sirota et al. (2014) and Johnson and Tubau (2013) for some further speculation]. Of further interest is exactly how these individual differences in facilitation are manifest in terms of differential thought processes that separate good Bayesian reasoning from bad.

Other recent research, for instance, is beginning to unearth exactly how different cognitive abilities inform different forms of reasoning (e.g., Del Missier et al., 2013). Elsewhere, De Neys and Bonnefon (2013) consider that cognitive individual differences may occur either early or late in the reasoning process. Their contention is that early individual divergences in the reasoning process may represent a more fundamental lack of formal knowledge, whilst later divergences may represent failures in appropriately applying knowledge. Given this hypothesis, individual differences in facilitation effects could be leveraged to signal the particular step in the Bayesian process on which a particular intervention exerts most benefit. For this type of approach to yield maximum insight, however, requires more than a slight shift in theoretical perspective; it will also require a reappraisal of some typical methodological practices used in the study of Bayesian reasoning.

Mandel (2014) succinctly notes several issues that have typified the archetypal methods used to study Bayesian reasoning, notably that of using word problems such as Eddy's (1982) mammography problem. Whilst the use of word problems can provide a convenient litmus test of one's capacity for Bayesian thought, they are often studied in ways that afford limited insight into reasoners' thinking. Two longstanding issues in particular can be identified that, if addressed, would complement attempts to understand how reasoners conduct the process of Bayesian reasoning, and how component abilities map onto this process.

Firstly, word problems predominantly focus on the endpoint of the judgment process, that is: whether someone produces the correct numerical estimate or not. We might conceive of the process of Bayesian judgment as akin to navigating a maze: there is usually one correct path to the exit, but several dead ends that one may arrive at before identifying the correct path. The process of Bayesian reasoning, for most people, may involve a similar process of cognitive tribulation before one reaches the point of arithmetic computation. Yet, by focusing on the endpoint we learn little about the journey. In doing so, research eschews potential opportunities to gain richer awareness into how interventions may change peoples' mental journey through the Bayesian maze, awareness that would further clarify the manner in which these interventions are effective. Future research, then, should look to study *how* reasoners reach their final Bayesian judgments, rather than simply what that final judgment is. One suggestion would be to make greater use of think-aloud protocols to identify the steps at which non-Bayesian deviations occur, and what such deviations entail. Whilst think-aloud paradigms are not without issue—verbalizing thoughts when reasoning can be cognitively challenging (Wilson, 1994); and the mere act of thinking aloud can reactively alter the reasoning process (e.g., Ericsson and Simon, 1998)—the process has previously yielded useful inferences into the types of thoughts underlying errors in Bayesian reasoning (De Neys and Glumicic, 2008). Potential procedural issues are also not without remedy. Although asking reasoners to think-aloud whilst solving more complex Bayesian word problems may prove overly-taxing for the average person, an alternative approach might see the Bayesian task broken down into component steps such as, for instance, information selection; information integration; and finally calculation (see Krynski and Tenenbaum, 2007, for a similar conceptualization). Reducing the overall task into component subtasks presented sequentially may reduce the overall burden of a think-aloud paradigm in this context, and more importantly maximize insight into the exact points in the Bayesian maze at which people deviate from the normative path, permitting more fine-grained interpretations. Varying the think-aloud procedure between subjects should also control for any concern regarding whether a think-aloud approach might actually alter how people would otherwise think about and reason through the task.

A second longstanding issue concerns how research often denotes participant estimates as "correct" (i.e., Bayesian) or "incorrect" (i.e., all other responses). Focusing on the accuracy of judgments alone may conceivably mean an indeterminate number of respondents are perhaps harshly categorized as poor Bayesian reasoners on account of failing to compute a strictly normatively accurate estimate. McNair and Feeney (2013), for instance, observed negligible levels of Bayesian responding on a mammography problem when only exactly arithmetically correct responses were accepted, yet consistently observed that a quarter of all responses fell within 5% of the correct estimate. Furthermore, the specific errors people produce offer potentially rich insights as to how the final judgment was conceived (e.g., Gigerenzer and Hoffrage, 1995); an overly conservative judgment connotes a very different thought process to a wildly inflated estimate. Future research may look to leverage Zhu and Gigerenzer's (2006) "write-aloud" procedure, as an example, which not only identifies a range of discrete errors—each characterized by different reasoning—but also precludes those who produce marginally incorrect estimates as being classified as de facto poor reasoners. Furthermore, rather than dichotomizing responses—which may give a diminished sense of an intervention's effectiveness—reporting *graded improvements* in accuracy (e.g., number of judgments within 5, 10, or 15% of the arithmetic estimate etc.) may also provide an altogether more rigorous evaluation of an intervention's capacity for facilitation.

Research on Bayesian facilitation continues to be productive, as evidenced by the recent upturn in research on individual differences in facilitation effects. Facilitating Bayesian reasoning, ultimately, requires an understanding of the "cognitive tools" people need in order to

make such judgments (Ayal and Beyth-Marom, 2014), and how these are applied when engaging in the mental process of Bayesian reasoning. What do people do when navigating the Bayesian maze? At what "step" in the process do deviations from the normative path occur, and are such errors predicted by particular cognitive limitations? The developing picture regarding cognitive capacities and Bayesian reasoning represents an ideal opportunity to more-closely address such questions, but in doing so research must do more to resist certain tendencies that have become somewhat ingrained into the study of Bayesian reasoning. Overcoming these status-quos stands to further elevate our understanding of "what works and why" in Bayesian facilitation through providing greater specifications of the cognitive minutiae involved in producing Bayesian judgments than is currently provided by existing theoretical accounts.

Future research should perhaps look to investigate how specific cognitive capacities relate to each component "step" in the Bayesian reasoning process, taking care to also specify the types of errors produced at each stage, and doing more to distinguish good reasoning and bad arithmetic. The use of more process-oriented methods, such as those considered earlier, can afford a much greater level of fidelity in achieving these goals, and will offer greater insight into what it means to "be Bayesian"—how reasoning progresses; and how, when, and why it sometimes falters. It follows that such research will allow for more targeted refinements in our understanding of what types of intervention strategies may apply best in facilitating better judgments in domains such as health, law, policy, and finance.

ACKNOWLEDGMENTS

I thank Wandi Bruine de Bruin and Barbara Summers for helpful comments on earlier drafts of this article. I also thank both reviewers for their insightful advice and constructive comments.

REFERENCES

Ayal, S., and Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgm. Decis. Making* 9, 226–242.

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychol. Bul. Rev.* 15, 284–289. doi: 10.3758/PBR.15.2.284

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol.* 381, 369–381. doi: 10.1002/acp.1460

Brase, G. L., Fiddick, L., and Harries, C. (2006). Participant recruitment methods and statistical reasoning performance. *Q. J. Exp Psychol.* 59, 965–976. doi: 10.1080/02724980543000132

Casscells, W., Schoenberger, A., and Grayboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *N. Engl. J. Med.* 299, 999–1000. doi: 10.1056/NEJM197811022991808

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

Del Missier, F., Mäntylä, T., Hansson, P., Bruine de Bruin, W., Parker, A., and Nilsson, L.-G. (2013). The multifold relationship between memory and decision making: an individual differences study. *J. Exp Psychol. Learn. Mem. Cogn.* 39, 1344–1364. doi: 10.1037/a0032379

De Neys, W., and Bonnefon, J. F. (2013). The whens and whys of individual differences in individual thinking biases. *Trends. Cogn. Sci.* 17, 172–178. doi: 10.1016/j.tics.2013.02.001

De Neys, W., and Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition* 106, 1248–1299. doi: 10.1016/j.cognition.2007.06.002

Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine," in *Judgment Under Uncertainty: Heuristics and Biases,* eds D. Kahneman, P. Slovic, and A. Tversky (New York, NY: Cambridge University Press), 249–267. doi: 10.1017/CBO9780511809477.019

Ericsson, K. A., and Simon, H. A. (1998). How to study thinking in everyday life: contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind Cult. Act.* 5, 178–186. doi: 10.1207/s15327884mca0503_3

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

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Hill, W. T., and Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *Q. J. Exp. Psychol.* 65, 2343–2368. doi: 10.1080/17470218.2012.687004

Johnson, E. D., and Tubau, E. (2013). Words, numbers, and numeracy: diminishing individual differences in Bayesian reasoning. *Learn. Individ. Dif.* 28, 34–40. doi: 10.1016/j.lindif.2013.09.004

Krynski, T. R., and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *J. Exp. Psychol. Gen.* 136, 430–450. doi: 10.1037/0096-3445.136.3.430

Lesage, E., Navarrete, G., and De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: the role of general cognitive resources. *Think. Reason.* 19, 27–53. doi: 10.1080/13546783.2012.713177

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Front. Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

McNair, S., and Feeney, A. (2013). When does information about causal structure improve statistical reasoning? *Q. J. Exp. Psychol.* 67, 625–645. doi: 10.1080/17470218.2013.821709

McNair, S., and Feeney, A. (in press). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychol. Bull. Rev.* doi: 10.3758/s13423-014-0645-y

Navarrete, G., Correia, R., and Froimovitch, D. (2014). Communicating risk in prenatal screening: the consequences of Bayesian misapprehension. *Front. Psychol. Cogn.* 5:1272. doi: 10.3389/fpsyg.2014.01272

Nobles, R., and Schiff, D. (2007). Misleading statistics within criminal trials. *Med. Sci. Law* 47, 7–10. doi: 10.1258/rsmmsl.47.1.7

Sirota, M., Juanchich, M., and Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychol. Bull. Rev.* doi: 10.3758/s13423-013-0464-6

Sloman, S. A., Over, D., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Process.* 91, 296–309. doi: 10.1016/S0749-5978(03)00021-9

Wilson, T. D. (1994). The proper protocol: validity and completeness of verbal reports. *Psychol. Sci.* 5, 249–252. doi: 10.1111/j.1467-9280.1994.tb00621.x

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: frequency or nested sets? *Exp. Psychol.* 50, 97–106. doi: 10.1026/1618-3169.50.2.97

Zhu, L., and Gigerenzer, G. (2006). Children can solve Bayesian problems: the role of representation in mental computation. *Cognition* 98. 287–308. doi: 10.1016/j.cognition.2004.12.003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 December 2014; accepted: 18 January 2015; published online: 06 February 2015.
*Citation: McNair SJ (2015) Beyond the status-quo: research on Bayesian reasoning must develop in both theory and method. Front. Psychol. **6**:97. doi: 10.3389/fpsyg.2015.00097*
This article was submitted to Cognition, a section of the journal Frontiers in Psychology.

frontiers in Psychology

PERSPECTIVE
published: 11 August 2015
doi: 10.3389/fpsyg.2015.01168

Rationality, the Bayesian standpoint, and the Monty-Hall problem

Jean Baratgin[1,2]*

[1] Laboratory CHArt (PARIS), Université Paris 8, Paris, France, [2] Institut Jean Nicod, Paris, France

The Monty-Hall Problem (*MHP*) has been used to argue against a subjectivist view of Bayesianism in two ways. First, psychologists have used it to illustrate that people do not revise their degrees of belief in line with experimenters' application of Bayes' rule. Second, philosophers view *MHP* and its two-player extension (MHP_2) as evidence that probabilities cannot be applied to single cases. Both arguments neglect the Bayesian standpoint, which requires that MHP_2 (studied here) be described in different terms than usually applied and that the initial set of possibilities be stable (i.e., a focusing situation). This article corrects these errors and reasserts the Bayesian standpoint; namely, that the subjective probability of an event is always conditional on a belief reviser's specific current state of knowledge.

Keywords: Bayesian standpoint, Monty-Hall problem with two players, probability revision, collider principle, single case probability

OPEN ACCESS

Edited by:
David R. Mandel,
Defence Research and Development Canada, Canada

Reviewed by:
Jan Sprenger,
Tilburg University, Netherlands
Peter Baumann,
Swarthmore College, USA

***Correspondence:**
Jean Baratgin,
Laboratory CHArt, Université Paris 8,
Site Paris-EPHE: 4–14 rue Ferrus,
75014 Paris, France
jean.baratgin@univ-paris8.fr

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *30 March 2015*
Accepted: *24 July 2015*
Published: *11 August 2015*

Citation:
Baratgin J (2015) Rationality, the Bayesian standpoint, and the Monty-Hall problem. Front. Psychol. 6:1168. doi: 10.3389/fpsyg.2015.01168

1. Introduction

In *the Monty Hall Problem* (*MHP*), you know that the car you want is behind one of three closed doors and a goat behind the other two doors. You choose a door and Monty (the host who knows where the car is) opens another door with a goat behind (as you know he can neither open your door nor a door with the car behind). After the host's action, would you rather stick to your original choice or switch to the remaining door?

MHP is a much-studied experimental paradigm investigating the inability of (naive and expert) people to revise their degrees of belief in a Bayesian manner (for a recent review see Tubau et al., 2015). Specific reformulations of format (natural frequencies, nested sets, visual representation, etc.) improving Bayesian performance have triggered some psychological debates on the underlying cognitive processes at play (for a recent analysis see Brase and Hill, 2015). Baratgin (2009) argues that these different formats facilitating Bayesian performance actually enhance the correct representation of the situation of revision in a stable universe, called the situation of *focusing* (Dubois and Prade, 1992, 1997) for which only Bayes' rule applies. The standard formulation of *MHP* prompts participants to form different representations of the situation of revision. However, when participants perceive the situation of focusing (for instance in a disambiguated version of *MHP* as in Baratgin and Politzer, 2010), they produce the Bayesian answer. Hence, participants cannot be considered as incoherent but only prone to an error induced by experimenters' presentation (Baratgin, 2009; Baratgin and Politzer, 2010).

MHP is also used as an argument against the notion of single-case probabilities. Moser and Mulder (1994) argued that there existed two opposite rational solutions: "sticking" for a *MHP* proposed as a one-shot problem and "switching" for a *MHP* cast in a frequentist context (i.e., when imagining a sufficiently large number of games). Horgan (1995) opposed this view making explicit the correct solution for the one shot *MHP* and showing that switching is the only correct solution to

both formulations. Baumann (2005, 2008) produced a new argument based on a generalization of *MHP*: *the Monty Hall Problem with two players* (MHP_2, see **Table 1**). In his view, although the two players share the same initial state of knowledge, they eventually form two different probability distributions. This point of view is opposed by Levy (2007) and by Sprenger (2010) who rightly argue that the two players do not necessarily share the same state of knowledge *throughout* the game in particular when their original choices differ. However, these authors do not explain the rationale of Baumann's mistake and do not explicitly define the causal structure of MHP_2[1].

This paper will address these questions. First, the solution to MHP_2 proposed as a one shot and its causal structure will be detailed. Then, explanations for the failure of researchers investigating MHP_2 will be advanced and related to the "bias" that conducts psychologists to wrongly conclude that participants' responses to *MHP* are of a non-Bayesian nature, that is, the *neglect of the Bayesian standpoint* (de Finetti, 1974).

2. Solving the Monty Hall Problem with Two Players

Let's consider the following variables that define the properties of the possible doors (D_1,D_2,D_3) in MHP_2: The three variables C (*The host's original choice of the door in which to place the car*), Y (*Your original choice of door*) and B (*Player B's original choice of door*). C, Y, and B can take any of the three values D_i (with $i \in \{1, 2, 3\}$), respectively noted from now on c_i, y_i, and b_i. The variable H (*the host's choice when opening a door*) is composed of the two complementary sub variables 'G' (*the host's revealing a goat*) and 'C' (*the host's revealing a car*). The sub variables 'G' and 'C' can take the three values D_i (with $i \in \{1, 2, 3\}$), respectively noted from now on 'g_i' and 'c_i'[2].

Following Walliser and Zwirn (2011), your beliefs before learning message 'g_3' assuming your initial choice is $D1$ (Stage 2) can be represented as a hierarchical dynamic probabilistic structure (see **Figure 1**). The layer 0 depicts the four possible strategies of the host, i.e., showing a goat behind D_2 or D_3 ('g_2' or 'g_3') or showing a car when the two players have originally chosen two different doors with goats behind ('c_2' or 'c_3'). Layer 1 corresponds to the three possible original choices of player B (b_1, b_2 or b_3). Layer 2 represents the original car placement choice of the host (c_1, c_2, or c_3). Layer 3 is your original choice (y_1). The probability distributions of the variables at the different layers are defined by the statement of MHP_2 with implicit and explicit hypotheses about the host's action and the players' preferences.

At Stage 4 you learn that the host will open a door with a goat behind. You know that (i) this door is either door D_2 or D_3 and (ii) the car is either behind your door D_1 or player B's originally chosen door. Hence you focus on the subset where 'g_2' or 'g_3' is true (the continuous lines in **Figure 1**). You are better off sticking to your initial choice D_1.

$$P(c_1|y_1\text{'}G\text{'}) = 3/7 > 2/7 = P(c_2|y_1\text{'}G\text{'}) = P(c_3|y_1\text{'}G\text{'}) \quad (1)$$

Second at Stage 5 the host opens door D_3 and reveals a goat behind. You focus on the subset where 'g_3' is true (the bold lines in **Figure 1**). This information combined with your original choice of door provides information about the door behind which Monty placed the car. You are better off switching to door D_2.

$$P(c_1|y_1\text{'}g_3\text{'}) = 3/7 < 4/7 = P(c_2|y_1\text{'}g_3\text{'}) \quad (2)$$

Finally at Stage 6 you learn what was player B's original choice. On the one hand, it can coincide with yours (b_1). Both players are then exactly in the same situation with the same common knowledge. MHP_2 amounts to *MHP*. Hence, you know that C is twice as likely to have the value c_2 as to have the value c_1. The best strategy is to switch from your original choice to the other closed door D_2.

$$P(c_1|y_1b_1\text{'}g_3\text{'}) = 1/3 < 2/3 = P(c_2|y_1b_1\text{'}g_3\text{'}) \quad (3)$$

On the other hand you may learn that player B's original choice is different from yours (b_2). In this case there is no best strategy and you are indifferent to sticking or switching.

$$P(c_1|y_1b_2\text{'}g_3\text{'}) = 1/2 = P(c_2|y_1b_2\text{'}g_3\text{'}) \quad (4)$$

3. The Collider Principle

Glymour (2001) was the first to identify the causal structure in *MHP* as a situation where two independent variables that mutually influence another variable are dependent conditional on the value of the variable they both affect. In MHP_2, the three independent variables Y, B, and C symmetrically influencing (colliding with) another variable H (common effect) actually appear dependent conditionally on the values of the variable H. Hence observing the value of H provides some information on the possible values of Y, B or C. In the same way, knowing the values of any couple of variables (C, H), (B, H), and (Y, H) provides some information about the values of couples (Y, B), (Y, C), and (B, C), respectively. Finally observing the values of triples (Y, C, H), (B, C, H), (Y, B, H), respectively determines the values of variables B, Y, and C. Solving MHP_2 as a one shot game relies on the latter triple (Y, B, H). It is easy when two variables are fixed to derive some qualitative predictions (Wellman and Henrion, 1993). For instance, MHP_2's solution supports a phenomenon of reversal decision resulting from this collider principle. On learning $H = $ 'g_3' given your original choice ($Y = y_1$) the likelihoods that B and C equal b_2 and c_2, respectively, are higher than the likelihoods that B and C equal b_1 and c_1, respectively. However, if in addition you learn that B equals b_1 then the outcome c_2 seems the more likely. However, if you learn that B equals b_2 then the probabilities for the car being behind either D_1 or D_2 are even.

[1]The term "causal" is missing in Baumann (2005). We find Horgan's terminology of "causal structure" in Levy (2007) with the vague definition of: "the set of conditions that ultimately explains why sticking and switching have the probabilities that they do" (Levy, 2007, p. 146). Finally, Sprenger (2010, p. 337) admits that "the place of causality in the 'causal structure' of a Monty Hall game remains obscure."

[2]We use here quotes for all sub-variables related to the host's actions during the game.

TABLE 1 | The six sequential stages of MHP_2.

Stages	Descriptions
Stage 1	The TV host shows to two players (players *A* and *B*) three identical doors (let them be D_1, D_2, and D_3) all equally likely, one hiding a car and the other two hiding goats. It is assumed that the host has no preference for a specific door when he initially places the car behind it and that both players prefer to win the car than a goat.[a] It is also assumed that the two players have a common initial state of knowledge and that no player has any preference for a particular door. The players fully grasp the six stages of MHP_2 and accept the implicit and explicit rules implied by its statement.
Stage 2	Each player picks a door and neither player is informed of the other player's choice. Let's assume for the sake of convenience that you are player *A* and you initially select door D_1.
Stage 3	The host, who knows where the car is, tells you: "In the case where player *B* has chosen the same door as you (here D_1), I will show you one door (out of the two other doors) behind which there is a goat." It is assumed that both players know that the host has no preference between the two remaining doors (D_2 and D_3) to show a goat should the car be behind D_1. Then the host continues: "In the case where player *B* has picked another door, I will always open the third door -chosen by neither player- even if the car is behind it." In this latter case when the host reveals a car, both players (you and player *B*) win and have no decision to make; the game stops.
Stage 4	The host says "I will open a door to reveal a goat" and then asks both players still ignorant of the other player's original choice: "To win the car should you stick to your original choice or switch to another door (as far as you are concerned door D_2 or door D_3)."
Stage 5	The host opens a door (for example D_3), reveals a goat and then asks both players again: "To win the car, should you stick to your original choice or switch to the other closed door (door D_2 in your case)?"
Stage 6	Each player reveals her or his original choice and must then decide knowing the other player's choice whether to stick to her/his door (D_1 in your case) or to switch door (D_2 in your case)[b].

[a] In the case where both players succeed in their door choice with the car, they each get a car. Hence, as noted by Sprenger (2010), there is no real competition between both players.
[b] This version of MHP_2 is derived from Baumann's version (Baumann, 2005). The transitional Stage 4 is not presented by Baumann but it interestingly draws a comparison with MHP where this information is not informative. We also added the Stage 6 to find again MHP in the situation where the two players have originally chosen the same door.

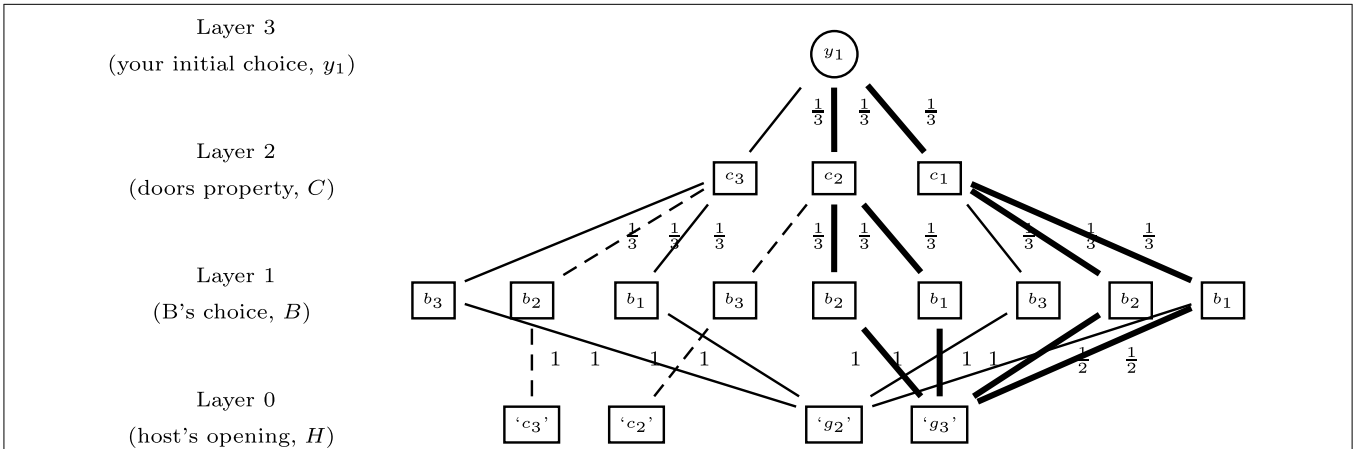

FIGURE 1 | The general tri-probabilistic structure of MHP_2 before learning message 'g_3' assuming your initial choice is D_1 ($Y = y_1$). The continuous lines correspond to the subset left after compiling information at Stage 4 and the bold lines to the subset left after compiling the information at Stage 5. Conversely the dashed lines represent the initial structure dropped out at Stage 4.

Recent studies have provided some evidence that "naive" adults and also children make correct qualitative predictions in collider principle situations when pairs of causal conditionals are explicitly presented (Ali et al., 2010, 2011). Precisely in *MHP*, participants perform better when the relation between the player's original choice and the host's strategy is explicit in conditional form (Macchi and Girotto, 1994, cited in Johnson-Laird et al., 1999). In the same way, when participants can construct a representation analogous to **Figure 1** for *MHP* using a graph or by means of physical handling, participants' performance improves significantly (Yamagishi, 2003; Baratgin and Politzer, 2010). Thus, it seems that when participants can infer the causal structure of *MHP* by physical or explanatory cues, they are able to solve *MHP* (Burns and Wieth, 2004; Chater and Oaksford, 2006).

4. The Neglect of the Bayesian Standpoint

De Finetti's subjective Bayesian standpoint proposes that individuals form two levels of knowledge (de Finetti, 1980; Baratgin and Politzer, in press):

- An elementary level of knowledge of an event E that is always conditioned on an individual's specific state of knowledge $\{H_0\}$ at this time. Furthermore, any event is actually a tri-event (the third value representing ignorance between true event and false event).
- A meta-level of knowledge concerning the degrees of belief of an individual. Here ignorance is specified, and refined, into degrees of belief. From an inferential point of view, your subjective probability of this event E at time t_0 is always *conditional on your current state of knowledge* $\{H_0\}$ [and should be written $P(E|H_0)$]. It is *coherent* if (i) it follows the axiom of additive probabilities[3] and (ii) when acquiring a new knowledge H, your probability also depends on this new knowledge $\{H_0H\}$ [and should be written $P(E|H_0H)$].

A person dismissing the Bayesian standpoint considers the probability of a single event as questionable as compared to a "frequentist" conception of probability. She takes the frequentist conception to be the "correct" comparative representation, and confines Bayesianism to just a set of *Bayesian techniques* (de Finetti, 1974). In the psychological literature this "bias" leads to two significant mistakes: (i) to the neglect of pragmatic constraints on the methodology (to understand H_0 and H); (ii) to the conclusion that people's behavior is "non-Bayesian," even when the behavior does not violate Bayesian coherence (Baratgin, 2002; Mandel, 2014a). In the analysis of MHP_2, this bias is characterized by inadequate terminology and interpretation of the revision situation.

4.1. The Use of an "Ambiguous Terminology"

For a subjective Bayesian, an event E always refers to a certain outcome in a single well-defined case (a unit in which the definition is unambiguous and complete) and cannot be used in a generic sense (such as a collection of "identical events"). There is no repetition of the same event but a succession of many distinct events, which can be different illustrations of the same phenomenon. In Moser and Mulder (1994), Baumann (2005), Levy (2007), and Baumann (2008), MHP_2 is presented in an *ambiguously termed* way (de Finetti, 1977/1981, p. 357). The variables are considered as trials of the same phenomenon without completely specifying them and their possible values. Every specific door corresponds to a generic door D that is characterized by two properties: having a car (C) or a goat (G) behind it. Every player's original door choice is analyzed by its correspondence with C and G. The host's door opening 'H' is characterized by the two sub-classes 'G' and 'C'. The players' final decisions to win the car are commingled and considered to pertain to the same classes of events "to stick," "to switch" or "nothing."

Following this *frequentist "jargon"* (de Finetti, 1979a,b), MHP_2 is analyzed as an observation of a repetitive problem where the different variables are interchangeable in function of the host's car placement. Instead of considering each player with specific states of knowledge relative to each stage of MHP_2 both players are assumed to have a *common knowledge* at each stage of the game. Their probabilities that there is a car behind one of the two remaining doors (after the door with a goat behind was opened) is 3/7 for the door originally chosen and 4/7 for the other door. Thus, imagining they made a different original choice, each door can be associated with two different probabilities (3/7 and 4/7) illustrating Bauman's paradox. Now, if we consider the specific knowledge of each player, the paradox disappears. In Stages 4 and 5, player B's probabilities on c_1 and c_2 are identical to your probabilities (relations 1–3) when his/her specific initial state knowledge is identical to yours (his/her original choice is b_1). Conversely when his/her original choice is b_2, his/her state of knowledge is different from yours and his/her probabilities correspond to different probabilities (relations 5 and 6):

$$P(c_1|b_2\text{'}G\text{'}) = P(c_3|b_2\text{'}G\text{'}) = 2/7 < 3/7 = P(c_2|b_2\text{'}G\text{'}) \quad (5)$$

$$P(c_1|b_2\text{'}g_3\text{'}) = 4/7 > 3/7 = P(c_2|b_2\text{'}g_3\text{'}) \quad (6)$$

However, player B's decisions are identical: sticking at Stage 4 and switching at Stage 5. At Stage 6, both players have an identical state of knowledge and probabilities (relation 4).

4.2. Neglect of the Situation of Focusing

MHP_2 illustrates that the situation of revision implied by the Bayesian standpoint is a process of *focusing* on a subset of the initial state of knowledge $\{H_0\}$ (de Finetti, 1957; Dubois and Prade, 1992, 1997). It is assumed that one object is selected from the universe and that a message releases information about it. A reference class different from the initial one is consequently considered by *focusing* attention on a given subset of the original set that complies with the information about the selected object. This is not a temporal revision process because the information 'g_3' just focuses on the selection of a particular posterior probability that was virtually available (among others) (see the bold lines of **Figure 1**). Yet participants in *MHP* seem to adopt (for pragmatic reasons) another representation of the revision situation, known as *updating* (Katsuno and Mendelzon, 1992; Walliser and Zwirn, 2002) in which, they infer from the message 'g_3' the information as "door D_3 have been removed," and conceive a new probability distribution consistent with this *new problem* (Baratgin and Politzer, 2007, 2010; Baratgin, 2009). In this representation there is obviously no collider effect because, in this *new problem with two doors*, the variables Y and H always remain independent after the information is provided by the host. Participants form a new distribution of probability P' for this new game[4]. Two typical analyses are consistent with this interpretation:

The *stick or switch response*: if you originally chose door D_1 and the host opens door D_3 with a goat behind, the worlds c_1 and c_2 are evenly close (in fact proportionally to their prior probabilities) to the invalidated world c_3. The weight of c_3

[3] See for example on this special research topic (Cruz et al., 2015; Evans et al., 2015; Mandel, 2015) and also (Politzer and Baratgin, in press).

[4] P' along the following process: (i) The worlds 'c_3' and 'g_2' are canceled and a simpler probabilistic structure composed of the two worlds (c_1, c_2) is obtained, (ii) The new distribution P' stems from a redistribution of the weights (the probabilities) of the removed worlds on the two remaining worlds.

is redistributed proportionally on c_1 and c_2. This is *MHP*'s solution in the updating context proposed by Dubois and Prade (1992).

$$\begin{aligned} P'(c_1|y_1) &= P(c_1|y_1) + 1/2P(c_3|y_1) = 1/2 \\ &= P(c_2|y_1) + 1/2P(c_3|y_1) = P'(c_2|y_1) \end{aligned} \quad (7)$$

It corresponds to the "equiprobability" solution given by nearly all participants to *MHP* but also by some experts in their analysis of *MHP* in a single isolated situation (Moser and Mulder, 1994) and of MHP_2 (Levy, 2007).

The *switch response*: The worlds c_3 and c_2 (the two doors not originally chosen by the player) are considered closer. The probability of the invalidated world c_3 is transferred to c_2 alone. This is *MHP*'s solution in the updating context proposed by Cross (2000).

$$\begin{aligned} P'(c_1|y_1) &= P(c_1|y_1) = 1/3 \text{ and} \\ P'(c_2|y_1) &= P(c_2|y_1) + P(c_3|y_1) = 2/3 \end{aligned} \quad (8)$$

This response is given by only few participants to *MHP* (see for review Baratgin, 2009). It corresponds to Moser and Mulder's explanation for *MHP*'s solution in a suitable long run of relevantly similar situations. To explain the "causal structure" of *MHP*, Levy (2007) proposed also a process in line with this updating interpretation. However, it is difficult here to support the "switch" response to MHP_2 with the symmetric role of the two players (Levy, 2007). Thus, the "stick or switch response" should be privileged to solve MHP_2 in an updating representation.

5. Conclusion

This paper describes the supposedly paradoxical solutions attributed to MHP_2 from the perspective of a thorough Bayesian standpoint perspective. It outlines the methodological care that one should take to comprehend the problem in relation to the single case terminology and the focusing context of revision. Not taking into account these features prevents one from fully grasping the probabilistic temporal dynamics of the problem and consequently the corresponding causal collider structure.

Psychologists who study subjective Bayesian reasoning should carefully formulate the statement without ambiguity and respect the Bayesian standpoint. This is also true especially for complex problems (such as the Sleeping Beauty problem Baratgin and Walliser, 2010; Mandel, 2014b) in which different solutions can be envisaged depending on the interpretations made by participants.

Acknowledgments

Financial support for this work was provided by a grant from the ANR Chorus 2011 (project BTAFDOC). The author thanks N. Cruz, G. Politzer, and B. Walliser for very helpful comments on a previous draft of this manuscript.

References

Ali, N., Chater, N., and Oaksford, M. (2011). The mental representation of causal conditional reasoning: mental models or causal models. *Cognition* 119, 403–418. doi: 10.1016/j.cognition.2011.02.005

Ali, N., Schlottmann, A., Shaw, A., Chater, N., and Oaksford, M. (2010). "Causal discounting and conditional reasoning in children," in *Cognition and Conditionals. Probability and Logic in Human Thinking*, eds M. Oaksford and N. Chater (New York, NY: Oxford University Press), 117–134.

Baratgin, J. (2002). Is the human mind definitely not bayesian? A review of the various arguments. *Curr. Psychol. Cogn.* 21, 653–682.

Baratgin, J. (2009). Updating our beliefs about inconsistency: the Monty-Hall case. *Math. Soc. Sci.* 57, 67–95. doi: 10.1016/j.mathsocsci.2008.08.006

Baratgin, J., and Politzer, G. (2007). The psychology of dynamic probability judgment: order effect, normative theories and experimental methodology. *Mind Soc.* 6, 53–66. doi: 10.1007/s11299-006-0025-z

Baratgin, J., and Politzer, G. (2010). Updating: a psychologically basic situation of probability revision. *Think. Reason.* 16, 253–287. doi: 10.1080/13546783.2010.519564

Baratgin, J., and Politzer, G. (in press). "Logic, probability and inference: a methodology for a new paradigm," in *Cognitive Unconscious and Human Rationality*, eds L. Macchi, M. Bagassi, and R. Viale (Cambridge, MA: MIT Press).

Baratgin, J., and Walliser, B. (2010). Sleeping beauty and the absent-minded driver. *Theory Decis.* 69, 489–496. doi: 10.1007/s11238-010-9215-6

Baumann, P. (2005). Three doors, two players, and single-case probabilities. *Am. Philos. Q.* 42, 71–79. Available online at: http://www.jstor.org/stable/20010183?seq=1#page_scan_tab_contents

Baumann, P. (2008). Single-case probabilities and the case of Monty Hall: Levy's view. *Synthese* 162, 265–273. doi: 10.1007/s11229-007-9185-6

Brase, G. L., and Hill, W. T. (2015). Good fences make for good neighbors but bad science: a review of what improves bayesian reasoning and why. *Front. Psychol.* 6:340. doi: 10.3389/fpsyg.2015.00340

Burns, B., and Wieth, M. (2004). The collider principle in causal reasoning: why the Monty Hall dilemma is so hard. *J. Exp. Psychol.* 133, 434–449. doi: 10.1037/0096-3445.133.3.434

Chater, N., and Oaksford, M. (2006). "Information sampling and adaptive cognition," in *Mental Mechanisms. Speculations on Human Causal Learning and Reasoning*, eds K. Fiedler and P. Juslin (Cambridge: Cambridge University Press), 210–236.

Cross, C. B. (2000). A characterization of imaging in terms of Popper functions. *Philos. Sci.* 67, 316–338. doi: 10.1086/392778

Cruz, N., Baratgin, J., Oaksford, M., and Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Front. Psychol.* 6:192. doi: 10.3389/fpsyg.2015.00192

de Finetti, B. (1957). L'informazione, il ragionamento, l'inconscio nei rapporti con la previsione. *L'industria* 2, 3–27.

de Finetti, B. (1974). Bayesianism: its unifying role for both the foundations and applications of statistics. *Int. Stat. Rev.* 42, 117–130.

de Finetti, B. (1977/1981). La probabilità: guardarsi dalle contraffazioni. *Scientia* 111, 255–281.

de Finetti, B. (1979a). Jargon-derived and underlying ambiguity in the field of probability. *Scientia* 114, 713–716.

de Finetti, B. (1979b). Probability and exchangeability from a subjective point of view. *Int. Stat. Rev.* 47, 129–135.

de Finetti, B. (1980). "Voice probabilità," in *Encyclopedia*, (Torino: Einaudi), 1146–1187.

Dubois, D., and Prade, H. (1992). Evidence, knowledge, and belief functions. *Int. J. Approx. Reason.* 6, 295–319. doi: 10.1016/0888-613X(92)90027-W

Dubois, D., and Prade, H. (1997). "Focusing vs. belief revision: a fundamental distinction when dealing with generic knowledge," in *Qualitative and Quantitative Practical Reasoning*, Vol. 1244 *of Lecture Notes in Computer*

Science, eds D. Gabbay, R. Kruse, A. Nonnengart, and H. Ohlbach (Berlin; Heidelberg: Springer), 96–107.

Evans, J. S. Thompson, V. A., and Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Front. Psychol.* 6:398. doi: 10.3389/fpsyg.2015.00398

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: The MIT Press.

Horgan, T. (1995). Let's make a deal. *Philos. Pap.* 24, 209–222. doi: 10.1080/05568649509506532

Johnson-Laird P. N., Legrenzi, P., Girotto, V., and Sonino-Legrenzi, M. S. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychol. Rev.* 106, 62–88. doi: 10.1037/0033-295X.106.1.62

Katsuno, A., and Mendelzon, A. (1992). "On the difference between updating a knowledge base and revising it," in *Belief Revision*, ed P. Gärdenfors (Cambridge: Cambridge University Press), 183–203.

Levy, K. (2007). Baumann on the Monty Hall problem and single-case probabilities. *Synthese* 158, 139–151. doi: 10.1007/s11229-006-9065-5

Macchi, L., and Girotto, V. (1994). "Probabilistic reasoning with conditional probabilities: the three boxes paradox," in *Paper presented at the Annual Meeting of the Society for Judgement and Decision Making.* (St. Louis, MO)

Mandel, D. R. (2014a). The psychology of bayesian reasoning. *Front. Psychol.* 5:1144. doi: 10.3389/fpsyg.2014.01144

Mandel, D. R. (2014b). Visual representation of rational belief revision: another look at the sleeping beauty problem. *Front. Psychol.* 5:1232. doi: 10.3389/fpsyg.2014.01232

Mandel, D. R. (2015). Instruction in information structuring improves bayesian judgment in intelligence analysts. *Front. Psychol.* 6:387. doi: 10.3389/fpsyg.2015.00387

Moser, P. K., and Mulder, D. H. (1994). Probability in rational decision-making. *Philos. Pap.* 23, 109–128. doi: 10.1080/05568649409506416

Politzer, G., and Baratgin, J. (in press). Deductive schemas with uncertain premises using qualitative probability expressions. *Think. Reason.* doi: 10.1080/13546783.2015.1052561

Sprenger, J. (2010). Probability, rational single-case decisions and the Monty Hall problem. *Synthese* 174, 331–340. doi: 10.1007/s11229-008-9455-y

Tubau, E., Aguilar-Lleyda, D., and Johnson, E. D. (2015). Reasoning and choice in the monty hall dilemma (mhd): implications for improving bayesian reasoning. *Front. Psychol.* 6:353. doi: 10.3389/fpsyg.2015.00353

Walliser, B., and Zwirn, D. (2002). Can Bayes rule be justified by cognitive rationality principles? *Theory Decis.* 53, 95–135. doi: 10.1023/A:1021227106744

Walliser, B., and Zwirn, D. (2011). Change rules for hierarchical beliefs. *Int. J. Approx. Reason.* 52, 166–183. doi: 10.1016/j.ijar.2009.11.005

Wellman, M., and Henrion, M. (1993). Explaining 'explaining away'. *IEEE Trans. Pattern Anal. Mach. Intell.* 15, 287–292.

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: frequency or nested sets? *Exp. Psychol.* 50, 97–106. doi: 10.1026/1618-3169.50.2.97

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

REVIEW
published: 31 March 2015
doi: 10.3389/fpsyg.2015.00353

Reasoning and choice in the Monty Hall Dilemma (MHD): implications for improving Bayesian reasoning

Elisabet Tubau[1,2]*, *David Aguilar-Lleyda*[1,2] and *Eric D. Johnson*[1,2]

[1] Departament de Psicologia Bàsica, Facultat de Psicologia, Universitat de Barcelona, Barcelona, Spain, [2] Research Institute for Brain, Cognition and Behavior, University of Barcelona, Barcelona, Spain

The Monty Hall Dilemma (MHD) is a two-step decision problem involving counterintuitive conditional probabilities. The first choice is made among three equally probable options, whereas the second choice takes place after the elimination of one of the non-selected options which does not hide the prize. Differing from most Bayesian problems, statistical information in the MHD has to be inferred, either by learning outcome probabilities or by reasoning from the presented sequence of events. This often leads to suboptimal decisions and erroneous probability judgments. Specifically, decision makers commonly develop a wrong intuition that final probabilities are equally distributed, together with a preference for their first choice. Several studies have shown that repeated practice enhances sensitivity to the different reward probabilities, but does not facilitate correct Bayesian reasoning. However, modest improvements in probability judgments have been observed after guided explanations. To explain these dissociations, the present review focuses on two types of causes producing the observed biases: Emotional-based choice biases and cognitive limitations in understanding probabilistic information. Among the latter, we identify a crucial cause for the universal difficulty in overcoming the equiprobability illusion: Incomplete representation of prior and conditional probabilities. We conclude that repeated practice and/or high incentives can be effective for overcoming choice biases, but promoting an adequate partitioning of possibilities seems to be necessary for overcoming cognitive illusions and improving Bayesian reasoning.

Keywords: Bayesian reasoning, Monty Hall Dilemma, choice biases, cognitive illusions, reflection

OPEN ACCESS

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
Sangeet Khemlani, Naval Research Laboratory, USA
Carlos Santamaria, Universidad de La Laguna, Spain

***Correspondence:**
Elisabet Tubau, Departament de Psicologia Bàsica, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron 171, 08035 Barcelona, Catalonia, Spain
etubau@ub.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *13 January 2015*
Accepted: *12 March 2015*
Published: *31 March 2015*

Citation:
Tubau E, Aguilar-Lleyda D and Johnson ED (2015) Reasoning and choice in the Monty Hall Dilemma (MHD): implications for improving Bayesian reasoning. Front. Psychol. 6:353. doi: 10.3389/fpsyg.2015.00353

Introduction

Bayesian reasoning has primarily been investigated in the context of imaginary scenarios, in which participants are required to derive a posterior probability (or a posterior ratio of natural frequencies) from explicit statistical information. An exception can be found in research with the Monty Hall Dilemma (MHD), where Bayesian reasoning has been studied with both imaginary scenarios and repeated practice. Differing from typical Bayesian problems, priors and conditional probabilities in the MHD have to be inferred, either by learning reward probabilities or by reasoning from the presented sequence of events. By reviewing the main difficulties and interventions for improving either choice or probabilistic judgments in the MHD, two different causes of failures are introduced: (1) emotional-based choice biases (switch aversion and/or the endowment effect), and (2) cognitive limitations in understanding and representing probabilities. We argue

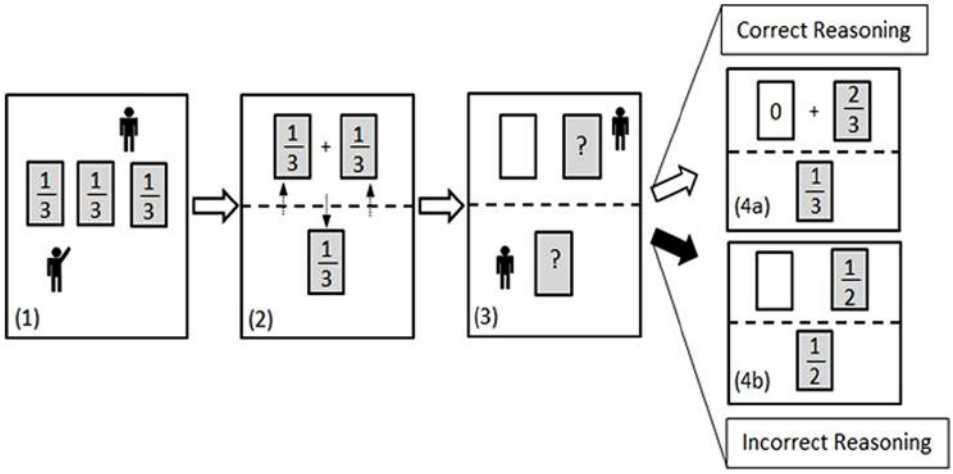

FIGURE 1 | Schematic representation of the MHD with the host (top) and player (bottom). (1) A player is presented three doors, each with an equal chance (1/3) of containing a prize and he chooses one of them. (2) Following the initial selection, the player now has one door with a 1/3 chance of having the prize. The host now has two doors with a total 2/3 chance of having the prize (1/3 + 1/3). (3) The host opens one of his two doors which does *not* contain the prize. The player is offered the choice to *stick* with his original selection or to *switch* to the unopened door held by the host. (4a) *Correct Reasoning*: Given that the opening of a non-rewarding door is obligatory, there still remains a 2/3 chance that the prize is on the "side" of the host, and a 1/3 chance that the prize is behind the player's originally chosen door. (4b) *Incorrect Reasoning*: A typical cognitive error is based on the illusion of equiprobability between the two remaining doors (see further explanation in the text).

that while the first cause produces illusions of control, regret, or distortions in the memory of past choice-outcome events, the second one promotes illusions of equiprobability and/or distortions in understanding the conditions of the game. The present review shows that both causes can independently and simultaneously bias choice and probabilistic judgments. Furthermore, whereas choice biases can be overcome by extended practice or by high incentives, overcoming the erroneous default intuition requires explicit instruction about the correct partitioning of probabilities. Implications for improving Bayesian reasoning are also discussed.

Understanding the MHD: From Intuition to Bayesian Reasoning

The MHD is a good example of a counterintuitive decision-making problem, considered to be "the most expressive example of *cognitive illusions* or *mental tunnels* in which even the finest and best-trained minds get trapped" (Piattelli-Palmarini, 1994; p. 161; cited by Krauss and Wang, 2003). In a first choice, a participant selects one of three possible options (i.e., doors), after being informed that only one hides a prize, and that the chances for each door are equal. Next, the host (or computer, in computer-based versions), who knows which door hides the prize, opens one non-rewarded door of the two remaining non-selected doors. The participant is then given a second, binary choice, which determines the final outcome of the game: They may either (a) stay with their initial selection [*stick*], or (b) swap their original selection for the other still closed door [*switch*]. The naïve reader would likely believe that each of the remaining two options has an equal probability of containing the prize, as often observed in the literature (i.e., Shimojo and Ichikawa, 1989; Franco-Watkins et al., 2003; Tubau and Alonso, 2003; De Neys and Verschueren, 2006; see also **Figure 1**). This common illusion has been attributed to a misapplication of the equiprobability principle (Falk, 1992; Johnson-Laird et al., 1999; Falk and Lann, 2008) due to the wrong intuition that, after the elimination of an option, all the chances have to be updated (Baratgin and Politzer, 2010). Specifically, the observation of two remaining options promotes the illusion that each of the final two options has a 50% chance of containing the prize. However, the elimination of an option (known by the host not to contain the prize) does not change the prior probability concerning the first choice. As shown in **Figure 1** and **Table 1**, the participant still has a 1/3 chance of having initially selected the prize and, therefore, in two out of three cases a decision to switch options will ultimately lead to a prize (a more formal explanation of probabilities in the MHD is introduced below).

Nevertheless, the final choice is generally neither fully coherent with the actual distribution of chances, nor with the misapplication of the equiprobability principle. A large majority of participants prefer to stick with the original choice (Granberg and Brown, 1995; Krauss and Wang, 2003), a tendency that has been related to an illusion of control (Lichtenstein and Slovic, 1971; Langer, 1975; Granberg and Dorr, 1998), or to a strategy to prevent future regret, which is more strongly perceived when losing after switching (Gilovich et al., 1995; Granberg and Brown, 1995; Petrocelli and Harris, 2011). Hence, the MHD motivates two different biases that work against the optimal solution: The equiprobability illusion and emotional-based choice biases. Both types of bias are difficult to overcome because the MHD presents an additional difficulty for most people: The need to distinguish a winning probability that has to be

TABLE 1 | Possibilities in the MHD: the probability of each door to be opened is conditioned on both the first choice and on the location of the prize.

Prize location	First choice	Probability to open door	Remaining door (after open)	Best choice
Door 1	**Door 1**	*P*(each door) = 0.5	Door 2 or Door 3	Stick
	Door 2	*P*(Door 3) = 1	**Door 1**	Switch
	Door 3	*P*(Door 2) = 1	**Door 1**	Switch
Door 2	Door 1	*P*(Door 3) = 1	**Door 2**	Switch
	Door 2	*P*(each door) = 0.5	Door 1 or Door 3	Stick
	Door 3	*P*(Door 1) = 1	**Door 2**	Switch
Door 3	Door 1	*P*(Door 2) = 1	**Door 3**	Switch
	Door 2	*P*(Door 1) = 1	**Door 3**	Switch
	Door 3	*P*(each door) = 0.5	Door 1 or Door 2	Stick

Notice that in 1 of 3 times the prize is behind the selected door and in 2 of 3 times the prize is behind the remaining door. Therefore, P(prize| stick) = 1/3 and P(prize| switch = 2/3 (in bold, location of the ace after opening one door).

updated (the one concerning the remaining door) from a winning probability that remains the same (the one concerning the first choice). Regarding this point, we claim that difficulties in overcoming illusions in the MHD are a consequence of a more primary cause: A biased representation of the prior probabilities. In Section "An Overlooked Failure: Incomplete Representation of Prior Probabilities" we review evidence supporting this claim.

From a Bayesian perspective, understanding the optimal solution in the MHD requires realizing that the elimination event is conditioned on both the first choice and on the location of the prize (Glymour, 2001; Burns and Wieth, 2004). Consider a scenario where the participant initially selects door 1. The conditional probability (likelihood) of eliminating, for example, door 3 after choosing door 1, depends on the hypothesis being considered (see also Falk and Lann, 2008). Specifically, given that the probability of revealing door 3 among the remaining two doors does not depend on the content of selected door 1 [$P(D_3|\ H_1) = P(D_3) = 1/2$], the posterior probability of such door containing the prize, conditioned to the elimination of door 3, is the same as its prior probability of containing the prize [$P(H_1|\ D_3) = P(H_1) = 1/3$]. In contrast, given that the probability of revealing door 3, conditioned to the prize being hidden in the remaining door 2, is doubled [$P(D_3|\ H_2) = 2P(D_3) = 1$], the posterior probability of door 2 hiding the prize, conditioned to the opening of door 3, also doubles [$P(H_2|\ D_3) = 2P(H_2) = 2/3$].

In other words, the conditions of the elimination have two main implications: (a) the winning probability for the selected door cannot change since it is conditioned to an unconditional event (it is certain that one of the non-selected doors is always null), and (b) the winning probability for the remaining door doubles, as the opening of a non-selected door is conditioned on the current location of the prize (see **Table 1**). In sum, understanding the MHD requires being able to distinguish conditional and unconditional events, or conditions in which probabilities have to be updated from conditions in which probabilities remain the same. In the following sections we review the difficulties found both in learning to choose optimally and in correct (explicit) Bayesian reasoning in the MHD in order to suggest causes and possible remediation.

Learning to Choose Optimally in the MHD

It is a well-grounded finding that both humans and non-human animals learn to optimize choices by adapting expectancies to the probability of forthcoming outcomes (Kahneman and Tversky, 1979). In repeated two-choice tasks, an increment in the probability of an optimal choice tends to follow the *matching law* (Herrnstein, 2000). Specifically, a matching between choice and reward probabilities is commonly observed, which is considered to be a consequence of a default adaptive strategy (West and Stanovich, 2003; Koehler and James, 2010). Nevertheless, sequential decision making tasks which include dependencies between choices can produce higher learning variability, and can lead to choices which deviate substantially from programmed reward probabilities (Herbranson and Wang, 2014).

Optimal choice in these more complex scenarios can be seen as arising from a Bayesian inference; that is, the probability of the outcome can be computed by combining its prior probability and the likelihood of the new observation. Alternatively, by repeating the decisional task, optimal choice preference can also develop through learning of either the most often rewarded final choice (i.e., switch in the MHD), or of the specific sequence of choices associated with the highest reward probability (e.g., "choose the leftmost option in the three-choice scenario, then switch in the two-choice decision"). The latter seems to explain pigeons' tendency to choose more optimally than humans in analogous MHD tasks (Herbranson and Schroeder, 2010; but see Mazur and Kahlbaugh, 2012 for similar results between species). In the case of humans, is repeated practice really useful for learning to choose optimally in the MHD? Furthermore, is this learning useful for improving correct Bayesian reasoning?

Since the earlier observations of Granberg and Brown (1995), several studies have shown an increase in switching rate after several repetitions of the MHD (Friedman, 1998; Granberg and Dorr, 1998; Franco-Watkins et al., 2003; Palacios-Huerta, 2003; Herbranson and Schroeder, 2010; Petrocelli and Harris, 2011; Mazur and Kahlbaugh, 2012; Klein et al., 2013; Saenen et al., 2014). However, in the absence of highly rewarding outcomes (Palacios-Huerta, 2003), a large majority of participants persist in the sub-optimal sticking strategy, switching in none or in only a few trials. As developed below, this impediment can be related to

a switching aversion and/or to an *endowment effect* (Kahneman et al., 1991). These emotional influences work against the discovery of the optimal choice by biasing the estimation of the winning probability of the first choice; that is, by inducing an illusion of control (Gilovich et al., 1995; Granberg and Brown, 1995), by biasing the memory of previous choice-outcome events (Petrocelli and Harris, 2011), and/or by preventing the accumulation of enough switching-winning experiences, as shown by a large number of participants in numerous studies.

Switch Aversion and the Endowment Effect

Similar to findings in other choice contexts (Landman, 1988), studies focusing on the MHD show that people report stronger regret when losing a prize by switching than by sticking (Gilovich et al., 1995; Granberg and Brown, 1995). Interestingly, Petrocelli and Harris (2011) observed that participants overestimated the trials in which they switched and lost, supporting the subjective experience that *switching and losing* is more aversive than *sticking and losing*. An increment of counterfactual thoughts associated with regret after switching and losing seemed to explain this distortion in memory (Petrocelli and Harris, 2011).

Not only do people find switching and losing highly aversive, they also appear to perceive switching and winning as less rewarding than sticking and winning (Franco-Watkins et al., 2003). In one of Franco-Watkins et al.'s (2003) experiments, participants played several rounds of the MHD after observing the choices and outcomes of a virtual participant in an analogous version of the game. Results showed that the switching rate of the participants was still below 50% even after observing that, in a rigged condition, switching produced 90% of winning trials, whereas the sticking rate was 100% after observing a player sticking and winning 90% of the trials (Franco-Watkins et al., 2003). Accordingly, the *win-stay, lose-switch* strategy shown in other probability learning tasks (e.g., Nowak and Sigmund, 1993) seems to be modulated by the previous choice (sticking or switching) in the MHD.

The switch aversion, or its complementary endowment effect—the tendency to attribute higher value to own options, even when compared to a slight more rewarding alternative (Kahneman et al., 1991)—has also been observed in variations of the MHD which include a larger number of doors (Stibel et al., 2009). That is, the endowment effect has been observed even in conditions where the difference between the final winning probabilities is much higher than in the standard three doors scenario (opening 8 of 9 remaining doors: Franco-Watkins et al., 2003; or opening 98 of 99 remaining doors: Stibel et al., 2009). In the mentioned experiment of Franco-Watkins et al. (2003), participants still preferred sticking with the initial choice even after observing the fictitious participant staying and losing in 90% of the trials (Franco-Watkins et al., 2003; 10C/3D condition). Stibel et al. (2009; Experiments 1 and 4) also found that between 30 and 50% of participants preferred the first choice after opening 98 of 99 remaining doors in one-shot game.

A marked tendency to stick with the first choice has also been observed in a condition in which the second choice was made between the first selection and *both* of the other two options, that is, without the elimination event and, hence, without the need to update probabilities (Morone and Fiore, 2008). As expected, the percentage of participants switching was significantly higher (across 10 trials, the overall switch rate was.58; 8 of 20 of participants had a switch rate higher than.7) compared to the standard MHD (the overall switch rate was.41; only 1 of 20 participants had a switch rate higher than.7). However, the percentage of participants with a switch rate below 0.5 was still not far away from the standard MHD (40 and 50% in "for dummies" and standard versions, respectively; Morone and Fiore, 2008), suggesting that switch aversion or the endowment effect work as attractors toward the non-optimal choice of sticking even in the MHD "for dummies."

Overcoming Choice Biases

Granberg and Dorr (1998), Tubau and Alonso (2003), and Stibel et al. (2009) attempted to reduce the endowment effect by eliminating the participants' first choice. This was accomplished by assigning participants one option among the initial three so that participants only had to choose between sticking and switching. Although the preference for switching was higher than in standard MHD conditions, about 50% of the participants still preferred the first, assigned choice (Tubau and Alonso, 2003). Furthermore, informal reports of the participants showed no improvement in correct Bayesian reasoning, including those participants who switched in most of the trials (Tubau and Alonso, 2003; see also Stibel et al., 2009). Typical comments of participants who finally became aware of the switching advantage believed that the computer program was biased in favor of switching but they expected the same winning probability for both choices (switching and sticking). It could be argued that such conditions hampered the motivation of the participants and, accordingly, their attention to the relevant contingencies was diminished. As observed in other tasks, being able to choose seems to be crucial to engage motivation (Leotti et al., 2010). But in the case of the MHD we have seen that the attraction to the first choice often prevents exploring the consequences of switching, making the discovery of the causes producing the switching advantage even more difficult.

On the other hand, it is well known that the perception of *two* remaining options in the final choice induces the misapplication of the equiprobability principle (Johnson-Laird et al., 1999; Falk and Lann, 2008). Hence, discovering the optimal choice in the MHD can be enhanced by changing the visual appearance of the final choice scenario or by manipulating the number of initial choices. For example, Howard et al. (2007) found higher switching rates in a condition in which all the boxes (closed and open) were visible compared to a condition in which the null options were removed. Increasing the area of the closed boxes also had a significant effect, although smaller than the number-of-boxes manipulation. Hence, the number of visible options seemed to be a relevant factor for promoting switching choices. Evidence of reasoning improvement was not reported but, based on other studies, it seems unlikely that the number-of-boxes manipulation had a significant effect on correct reasoning. In a one-shot scenario, Stibel et al. (2009) showed that, among the participants choosing to switch, probability judgments matched the equiprobability intuition, even in the condition in which 98 of the

remaining 99 options were removed! (see also Franco-Watkins et al., 2003).

In addition to the interventions introduced above, increasing incentives (Friedman, 1998; Palacios-Huerta, 2003), or enhancing collaborative playing (Palacios-Huerta, 2003) also seem to be effective for overcoming choice biases in the MHD, at least for some participants. It is worth noting that the most effective intervention appears to be the manipulation of incentives (Palacios-Huerta, 2003), supporting the emotional source of the choice biases observed in the MHD. Unfortunately, none of these latter studies reported probabilistic judgments of the participants. However, based on the results of Stibel et al. (2009), who also used money as a reward, an increment in the amount of gain does not seem to be effective for improving Bayesian reasoning. In the next section we review in more detail the relationship between choice and reasoning improvement in the MHD, as well as possible explanations for the observed dissociation.

Dissociating Choice from Reasoning

None of the MHD studies assessing the accuracy of probabilistic judgments after several repetitions have observed improvement of correct explicit Bayesian reasoning (Franco-Watkins et al., 2003; Klein et al., 2013; Saenen et al., 2014). In the best case, participants who, following practice, report that switching is more advantageous, tend to switch more often (Tubau and Alonso, 2003), but they are typically unable to explain the reason for that advantage (see also Klein et al., 2013).

It could be argued that the null effect of practice for enhancing understanding the probabilistic structure of the MHD is due to the small amount of practice (commonly less than 50 repetitions). Nevertheless, a larger number of trials appear insufficient for maximizing optimal choice (Herbranson and Schroeder, 2010; Klein et al., 2013; Saenen et al., 2014) or for enhancing correct Bayesian reasoning (Klein et al., 2013; Saenen et al., 2014). For example, after about 250 repetitions of the MHD, only one participant out of 17 seemed to correctly explain the optimal strategy: "First, I clicked on a random box. After one of the boxes disappeared, I clicked on the third box" (Klein et al., 2013), but even this was without clear evidence of having understood the *cause* of the switching advantage. Saenen et al. (2014) analyzed the accuracy of probability judgments in different moments during 100 repetitions of the MHD and found no evidence of improvement at any stage of practice. It is worth noting that Saenen et al. (2014) gave continuous feedback and, in one of the groups, feedback explicitly related winning and losing to each choice (sticking and switching). Although explicit feedback increased frequency of switching, it was not helpful for improving explicit probabilistic judgments.

Accordingly, studies centered on the effect of practice with the MHD suggest that knowledge acquired by learning the different winning probabilities does not lead to better comprehension of the MHD. More specifically, practice seems to facilitate the overcoming of initial choice biases, but does not facilitate an understanding of *why* initial choice tendencies are not optimal. Supporting this claim, significant increments in optimal choice in the MHD have been observed even without explicitly noticing its advantage, although the general tendency to choose optimally (switch) is much weaker than when noticing the switching advantage (Tubau and Alonso, 2003; Klein et al., 2013). In addition to the initial strong bias to avoid switching, these results suggest the involvement of associative mechanisms similar to the ones reported in studies with other non-human animals (Herbranson and Schroeder, 2010; Mazur and Kahlbaugh, 2012; Klein et al., 2013). Associative mechanisms can explain the observed implicit learning of the switching advantage. Nevertheless, without awareness of the rules and effortful control to apply them, they seem to be insufficient to overcome initial choice biases (see Stocco and Fum, 2008, for similar conclusion in other choice tasks).

In line with the associative account introduced to explain the observed dissociation between reasoning and choice, Stibel et al. (2009) concluded that evidential strength, on which choices are based, is sensitive to the evidence provided by alternative hypotheses, but explicit probability judgments are typically less sensitive to slight or apparent changes in support strength (see also Tversky and Koehler, 1994). Accordingly, variables affecting the increment of optimal choice, as for example the increment in the number of non-chosen options, produce an increment in evidence strength for the alternative hypothesis (switch in the MHD) without affecting the corresponding probabilistic judgment (Stibel et al., 2009). Similarly, the effect of repeated practice with the MHD enhances the realization that the proportion of winnings by switching is higher than winnings by sticking, which affects the evidence strength of the final choices. Nevertheless, all these interventions remain insufficient for overcoming the equiprobability illusion, which continues to bias explicit probabilistic judgments.

Enhancing Probabilistic Reasoning in the MHD

Based on the reviewed evidence, repeated practice and/or higher incentives have a moderate effect on increasing the probability to choose optimally, but it is not useful for enhancing the understanding of the causes of the switching advantage, namely, the prior, conditional, and posterior probabilities involved in the MHD. This section reviews the utility of interventions more directly aimed at improving explicit Bayesian reasoning.

Explaining Possibilities: Mental Models and the Perspective Effect

The information presented in the text of the problem affects the building of the mental models on which judgments and decisions are based (Legrenzi et al., 1993; Johnson-Laird et al., 1999). In the case of the MHD, different manipulations have been shown to affect reasoning and/or choice. As previously introduced, if instead of the standard dilemma, participants are offered a choice between the selected door and both of the remaining two doors ("for dummies" version in Morone and Fiore, 2008), the tendency to switch increases. It is well documented that decision makers create mental models based on the number of options being presented (Johnson-Laird et al., 1999). If one of the three options is removed, only two models are taken into account: One in which

the prize is behind the selected door, and one in which the prize is behind the remaining door (see also Johnson-Laird et al., 1999; Franco-Watkins et al., 2003). Nevertheless, presenting a more transparent MHD does not imply developing a more complete representation, as many individuals have trouble understanding the prior probabilities (Tubau and Alonso, 2003; Tubau, 2008; see below).

The interventions which have been demonstrated to be the most effective for improving correct reasoning in the MHD explicitly request the reasoner to imagine the different possibilities from the different perspectives of both the contestant and the host (Krauss and Wang, 2003; Tubau and Alonso, 2003; Tubau, 2008). For example, using a diagram, Krauss and Wang (2003) presented three closed doors, one representing the selection of a hypothetical contestant. To enhance the representation of the different possibilities from each perspective, participants were asked to imagine being the host of the game who is opening a null door between the two non-selected doors. The percentage of correct justifications of the switching advantage, from the contestant's point of view, increased from 3% in the standard MHD to 39% in this new presentation (50% correctly noticed the advantage of switching; Krauss and Wang, 2003). Given that participants did not perform the initial choice, it could be argued that the benefit of the intervention was in part a consequence of eliminating the difficulty in overcoming initial choice biases (see Switch Aversion and the Endowment Effect). However, the effectiveness of the perspective manipulation was also observed in an experienced adversary game context, regardless of the role of the participant (Tubau and Alonso, 2003).

More directly, Tubau and Alonso (2003) asked participants to represent the different possibilities from both perspectives. In their third experiment, participants were presented an imaginary card game between two adversaries: The decision maker selecting a card among three (one ace and two other cards), and the informant keeping the other two. Analogous to the host of the MHD, the informant always showed a non-ace card after the decision-maker's selection. In one experimental condition, participants had to state the possibilities of each player having the ace, and then estimate each player's likelihood of winning, as well as provide a justification for the perceived best strategy (switching, sticking, or no preference). This condition was compared to the same adversary version, but without the requirement of representing the possibilities, as well as to the standard MHD. Percentage of correct justifications for the switch response were 0% in the standard MHD, 25% in the adversary version without explicit representation, and 60% in the adversary version with explicit representation of possibilities. In sum, encouraging a shift between perspectives seems to be an effective intervention to enhance the building of more complete mental models of the different possible locations of the prize, as well as improved awareness of which options can be eliminated and why. Support for this proposal can also be found in Tor and Bazerman (2003) who, based on protocol analyses in different competitive games, concluded that the main difficulty in competitive contexts is to consider the decisions of others and the rules of the game (the constraints of the host in the MHD).

Enhancing Correct Probabilistic Judgments: The Role of Natural Frequencies

Another widely discussed facilitator of Bayesian reasoning performance is to present and request problem information as natural frequencies (Gigerenzer and Hoffrage, 1995; Girotto and Gonzalez, 2001; Johnson and Tubau, 2013). Although disagreement persists regarding the specific mechanisms involved in processing natural frequencies (e.g., Gigerenzer, 1994; Barbey and Sloman, 2007), presenting and requesting information in a similar frequency format is also known to facilitate reasoning in the MHD.

For example, Krauss and Wang (2003; Experiment 3) compared the utility of an intervention based on a simplified representation of only three arrangements (similar to first three possibilities in **Table 1**) with a more complete representation of six arrangements (mental model representation from Johnson-Laird et al., 1999; similar to the diagram shown in **Figure 2**, but including the complete representation of each possibility instead of the frequency information). Results showed that the three-arrangements version promoted more correct responses. The benefit of the simplified representation was interpreted as a consequence of its higher resemblance to a natural frequency format (Krauss and Wang, 2003). However, it is not clear which words and numbers were included in the question requiring the probability judgment. As shown in other Bayesian problems, the match between the text of the problem and the text of the question has a significant effect on the responses (Girotto and Gonzalez, 2001; Ayal and Beyth-Marom, 2014). If the question was the same as in Kraus and Wang's Experiment 2, then there would be a better match between the question (__ out of 3) and the simplified representation (three arrangements) than between the question and the complete version (six models). So, it could be the case that the more complete representation was

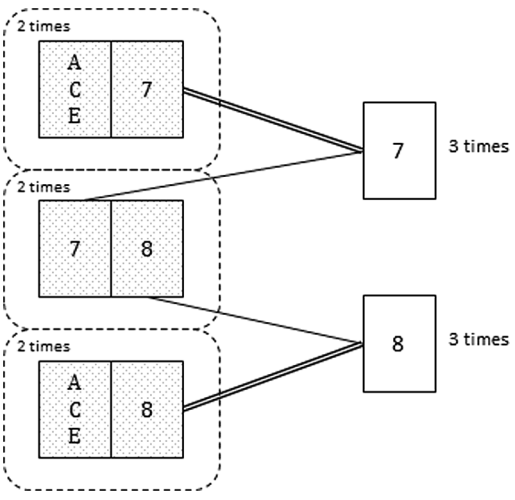

FIGURE 2 | Card shown by the informant (analogous to the host in the MHD) in six hypothetical repetitions of the game. Notice that among the three times that the informant shows the 7 (or the 8) he hides the ACE twice (adapted from Tubau, 2008).

less effective due to the additional steps needed to transform presented information into the form requested in the question.

Related to the previous hypothesis, in Tubau (2008; Experiments 1A,B) two explanations of an analogous MHD card game were compared: In the *concrete frequency* version, the explanation referred to a specific simulation of six games, analogous to the mental models representation (i.e., in the two cases in which John has the ace and the 7, he will show the 7; in the two cases in which John has the ace and the 8, he will show the 8; and in the two cases in which John has the 7 and the 8, he will show the 7 once and the 8 once; see **Figure 2**). In the *relative frequency* version, less precise verbal quantifiers were used (i.e., if John has the ace and the 7, he will *always* show the 7; if John has the ace and the 8, he will *always* show the 8, and if he has the 7 and the 8 he will show the 7 *half of the times* and the 8 *half of the times*). Each version was presented with and without a diagram similar to the one presented in **Figure 2**. Results showed a significant effect of statistical format (concrete frequencies enhanced performance compared to abstract quantifiers), but no effect was found for the visual diagram. Hence, results supported the Krauss and Wang (2003) and Tubau and Alonso (2003) conclusion regarding the need to build models (possibilities) from both perspectives in a way which facilitates the computation of the respective winning frequencies. As shown in these studies, the highest benefit is observed when participants are externally guided during both the presentation of the problem and via the formulation of the question. Furthermore, and similarly to other Bayesian reasoning problems, the closer the match between the numerical format included in the explanation and the required numerical expression, the higher the benefit (Girotto and Gonzalez, 2001; Ayal and Beyth-Marom, 2014).

Explaining Causal Relations: Competition Scenarios

According to the studies reviewed so far, probabilistic reasoning in the MHD can be improved through interventions that facilitate building a more complete representation of the different possibilities, or by prompting the required numerical expression in the format of the requested probabilistic judgment (i.e., ___ out of 3). Nevertheless, the extent to which any corresponding improvement indicates a complete understanding of both prior probabilities and the consequences of the elimination's conditions, (as opposed to simply being a consequence of a match between representations), remains unclear. As developed in Section "Understanding the MHD: From Intuition to Bayesian Reasoning," understanding the MHD implies understanding that, after the elimination of an option conditioned to the location of the prize, the winning probability of the first choice remains invariant, whereas the winning probability of the remaining option increases twofold.

Related to the comprehension of the elimination's constraints, a different and interesting approach to improve reasoning in the MHD was developed by Burns and Wieth (2004). Similarly to Glymour (2001), Burns and Wieth (2004) attributed the main cause of failed understanding of the MHD to a failure in understanding the causal structure which produces the switching advantage (see also Krynski and Tenenbaum, 2007, in other Bayesian scenarios). From this perspective, the fact that two independent causes (initial choice and location of the prize) collide on a common effect (the opening of one of the non-selected doors; see **Table 1**) might explain why the MHD is so hard. Based on this assumption, Burns and Wieth (2004) hypothesized that a context more clearly presenting the causes that determine the elimination of an option would enhance reasoning. Supporting this hypothesis, Burns and Wieth (2004) found better performance in analogous MHD competition scenarios (i.e., a competition among three boxers in which only one was the best). However, even in the best conditions of the competition context, only about 50% of the participants selected the optimal (switch) choice and less than 20% of participants were able to express the correct posterior winning probabilities. These results suggest that making more salient the causal conditions that determine the elimination event, or a better knowledge of the rules of the game (Tor and Bazerman, 2003), are also insufficient for a large number of participants to understand the MHD. It is worth noting that clear causal structures seem to primarily benefit higher numerate reasoners in other Bayesian problems (McNair and Feeney, 2014). In the case of the MHD, in addition to the just reviewed difficulties, we suggest that this limitation is also due to a failure in representing the prior probabilities.

An Overlooked Failure: Incomplete Representation of Prior Probabilities

How people represent the prior probabilities in the MHD has been rarely investigated. In most studies it is assumed that people have an accurate representation of the different probabilities before the elimination event that is, before inducing the equiprobability illusion. However, with the exception of the prior winning probability for the first choice, prior probabilities in the MHD are not necessarily obvious. Representing the winning probability of the initial choice is easy given the transparent correspondence between the initial information, three doors, and one prize, and the correct ratio 1 of 3 chances to win. However, representing the winning probability of the set including the two remaining doors might present a conflict between these two non-selected doors and the three initial doors. In fact, it has been observed that only about 50% of undergraduates understand that the chance of the non-selected options (held by the host or informant in the card game) hiding the ace is 2 of 3, with a common response instead being 1 of 2 (Tubau and Alonso, 2003; Tubau, 2008). Still more difficult is understanding (or expressing) that, among the non-selected options, at least one is null. Only 25% of participants were able to correctly answer the question: "What is the probability that, among the non-selected cards, at least one is not the ace?" (Tubau, 2008). Hence, although most participants are able to represent, in a diagram, the different possible locations of the prize (Tubau and Alonso, 2003), many have difficulties expressing the corresponding probabilities (Tubau and Alonso, 2003; Tubau, 2008).

Weak representation of uncertain information causes vulnerability to biases and/or to conservative behavior (van der Pligt, 1998). Similarly, we argue that one of the consequences of the

incomplete comprehension of prior probabilities in the MHD is the vulnerability to the equiprobability illusion. This, together with the choice biases discussed above, promotes the final decision to stick. In particular, susceptibility to the illusion is caused by a weak representation of the facts that: (a) the non-selected doors will hide the prize 2 out of 3 times, (b) among the non-selected doors it is certain that at least one is null, and (c) this null option will always be eliminated. Furthermore, without an adequate representation of the prior probabilities, the perspective manipulation commented above has no effect (e.g., in the adversary version without the explicit representation manipulation in Tubau and Alonso, 2003). Accordingly, being able to understand the elimination's conditions (the constraints imposed on the host or on the computer), which is crucial for correct Bayesian reasoning in the MHD (Krauss and Wang, 2003; Burns and Wieth, 2004), cannot be useful without an accurate representation of the prior probabilities. It is worth noting that the most effective intervention in Krauss and Wang (2003) was the one prompting reasoners to imagine themselves opening one of the doors according to the elimination's conditions (perspective effect), together with the requirement to express the answer as a ratio of frequencies: The number of times, out of 3, in which the prize would be behind the contestant's door. That is, the one promoting the representation of the initial possible locations of the prize.

In sum, a large number of the undergraduates that participate in the MHD experiments do not have adequate knowledge to understand and/or represent prior and conditional probabilities in the MHD (Tubau and Alonso, 2003; Tubau, 2008; see also Brase et al., 2006, for similar claim in the context of other probabilistic reasoning tasks). Therefore, when interpreting the data in the literature, it is important to take into account these limitations. A more complete comprehension of the psychology of the MHD would require the consideration of specific knowledge or skills as mediators of performance.

Understanding Reasoning Failures in the MHD: A Theoretical Analyses

Although not without critics (for a recent review see Evans and Stanovich, 2013), most current thinking theories share a dual-systems or dual-processing approach. In essence, dual thinking theories consider that effortless, intuitive thinking processes occasionally lead to erroneous or *suboptimal* responses, unless more effortful, analytical reasoning processes intervene to override an initially biased tendency (Evans, 2010; Kahneman, 2011; Stanovich, 2011). Some of the factors that determine the success of effortful reasoning include: Adequate cognitive resources, specific knowledge related to the task, confidence in the intuitive response, and thinking dispositions (engagement or *laziness* of the *reflective* mind). Specifically, Stanovich (2009) suggested that the reasoning system can be understood as including two different "minds": the algorithmic, which *controls* the running of specific reasoning procedures, and the reflective, which *decides* which reasoning algorithm to apply and/or whether or not to invest more effort into the task. Therefore, according to this proposal, overriding an erroneous response produced by the autonomous mind (Stanovich, 2009) might fail due to lack of resources and/or knowledge to run specific procedures (a failure of the algorithmic mind) and/or due to weak disposition to implement a needed procedure or to review an initial response (a failure of the reflective mind).

Applying this distinction to the MHD, would the frequent but wrong application of the equiprobability principle be a failure of the algorithmic mind? Or would it be consequence of a *lazy* reflective mind? As commented in Section "An Overlooked Failure: Incomplete Representation of Prior Probabilities," a large number of participants do not have adequate knowledge to correctly represent the prior and conditional probabilities in the MHD (e.g., the probability of the set of non-chosen doors containing the ace; the probability of one of the non-chosen doors being empty; Tubau and Alonso, 2003; Tubau, 2008). For these participants, explicit explanations of the different possibilities during the game had a weak effect on correct reasoning, compared to that observed with higher numerate participants (Tubau, 2008). In addition to a lack of specific knowledge, reasoning in the MHD has been also impaired when the reasoning resources (working memory) were compromised by a secondary task (De Neys and Verschueren, 2006), supporting the relevance of the algorithmic mind for correct reasoning. Nevertheless, it is a common finding that the MHD remains obscure even for high numerate individuals (Girotto and Gonzalez, 2005) or for participants with high working memory span (De Neys and Verschueren, 2006).

Regarding the role of the reflective mind in the MHD, there is no direct evidence of a relation between reflective thinking ability and performance in the MHD. Based on the general finding of strong difficulties in overcoming the equiprobability bias, even for individuals with more education (Girotto and Gonzalez, 2005) or higher working memory span (De Neys and Verschueren, 2006), we anticipate that the relation between reflective thinking capacity and correct reasoning in the MHD would be small or non-existent. It is possible that this relation might emerge if additional relevant information were provided (e.g., explicit representation of the different possibilities), as observed for participants higher in numeracy (Tubau, 2008). But, without this facilitation, weakness of the reflective mind on its own is unlikely to be the main cause of reasoning failures in the MHD.

If people high in cognitive reflection fail to review the erroneous default intuition it may be due to either an absence of the relevant triggering conditions for reflection, or to the absence of adequate knowledge to replace the erroneous default intuition with the correct model of the task (due, for example, to a biased representation of prior probabilities; see Section "An Overlooked Failure: Incomplete Representation of Prior Probabilities"). One of the relevant triggering conditions for reflection is the detection of conflicting beliefs, which tends to reduce confidence in the correctness of the response (Thompson, 2009; De Neys, 2014). In the case of the MHD, experience with the game can produce two different types of conflict: (1) Conflict between correct representation of prior probabilities and the elimination's conditions and the subsequent equiprobability intuition, and (2) Conflict between the default equiprobable intuition and the experienced

switching advantage. None of the reviewed studies have reported confidence measures or other measures of conflict detection. Nevertheless, based on previous findings showing incomplete prior representation and/or the formation of the wrong belief that, after the elimination of an option, a probability update is needed (Baratgin and Politzer, 2010), we anticipate that no conflict (1) would be detected, however, this would be an interesting question to follow up in future studies.

Related to previous conflict (2), there is evidence that noticing it does not improve the chances to override the default intuition (Tubau and Alonso, 2003; Klein et al., 2013; Saenen et al., 2014). For example, in Tubau and Alonso (2003), participants who noticed the conflict between the equiprobability intuition and the switching advantage seemed to solve this contradiction by creating a new explanation (in terms on an anomaly in the computer program). That is, the equiprobability intuition seemed to be accompanied by such a strong feeling of rightness (e.g., Thompson, 2009) that the observation of a discrepancy would have been associated with exception (anomalous program) rather than to a conflict to be solved. Furthermore, if some form of conflict were detected, the biased representation of the underlying probabilistic structure for most participants (see An Overlooked Failure: Incomplete Representation of Prior Probabilities), together with the direct perception of *two* final, *initially equal*, doors would have likely prevented finding the correct solution. In this sense, reasoning failures in the MHD could be attributed to automatic processes which build a particularly *vivid* default mental model of the task, and correspondingly strong justification of its correctness, rather than to a *weakness* of the reflective mind *per se*.

Implications for Enhancing Bayesian Reasoning

As commented above, participants noticing the switching advantage in the repeated MHD solved the contradiction with the default intuition by building an alternative explanation able to preserve it. This suggests that the *reflective mind* might indeed notice certain conflicting information (conflict 2 in previous section), but the relevant information needed to correct the error in the default intuition (i.e., correct representation of prior probabilities and the elimination's conditions) is either not available or ignored. Accordingly, the efficacy of interventions aimed at improving Bayesian reasoning in the MHD would depend on the available reasoner skills and/or external hints which enhance the building of a more complete representation of the task. According to the present review, the interventions that have been shown to be the most effective are the ones promoting a different partition of the probability space (Krauss and Wang, 2003; Tubau and Alonso, 2003; Tubau, 2008). Instead of modeling the winning probability of *each of the three options* separately [P(each option) = 1/3], understanding the MHD requires modeling the winning probability of each *set of possibilities* corresponding to each actor [i.e., P(contestant) = 1/3; P(host) = 2/3]. Notice that with this representation, and with the additional knowledge that the host for sure has at least one null option that must be shown, no other computation is needed (see **Table 2**).

TABLE 2 | Main beliefs and biases affecting reasoning and choice in the MHD both before and after the elimination of an option.

Incorrect reasoning and choice
Before the elimination of an option
Correct application of the equiprobability principle: ***Three equal* options** (frequently, together with incorrect or incomplete representation of the possibilities related to the set including the other two options)
After the elimination of a null option
Reasoning based on cognitive biases (incorrect comprehension of priors and/or the elimination's conditions; incorrect application of the equiprobability principle): ***Chances are equal for switch and stick* (1/2 each)**
Choice based on emotional biases (switch aversion; endowment effect; illusion of control): ***Select stick*** (consider switching in case of *bizarre* or *unexplainable* observation of switch advantage)
Correct reasoning and choice
Before the elimination of an option
Correct partition of the probability space: ***Two unequal* sets of possibilities**
Chances for the selected option: 1/3
Chances for the other *two* options: 2/3 (and a null option *for sure*)
After the *obligatory* elimination of a null option
Correct comprehension of the elimination's conditions
Chances for the selected option: 1/3
Chances for the other option: 2/3 (the null option was *predicted*)
Reasoning: *Chances are higher for switch (2/3) than for stick (1/3)*
Choice: *Switch is a better option than stick*

In sum, as observed in other Bayesian problems, the correct partition of the problem space of probabilities or corresponding set–subset structure is crucial for correct reasoning (Johnson-Laird et al., 1999; Barbey and Sloman, 2007). As also shown in other Bayesian problems, the use of natural frequencies can facilitate the comprehension of the MHD (Krauss and Wang, 2003; Tubau and Alonso, 2003; Tubau, 2008). This seems particularly relevant in case of lower numerate reasoners, who would require a simulation of the partitioned probabilities by simulating several repetitions of the game (Tubau, 2008). But, in general, reviewed findings in the MHD suggest that the accuracy of explicit Bayesian reasoning depends on the accuracy of the underlying partitions of the probability space included in the mental model of the task.

Conclusion

The strong counterintuitivity of the MHD has intrigued people for decades. What is it about the MHD that makes it so hard for people to know that switching is the best course of action to win the prize? And on top of that, what is it that generates such strong disbelief even if it is realized that switching is better? Assuming the random assignation of the prize, it is clear that, in the initial stage of the game, most people would correctly assign to each alternative the same probability of hiding the

prize. It is after the first choice is already made and the second choice to stick or switch is offered that the dilemma develops. The trouble starts with the initially built representation of the task upon which this second decision is based. On the one hand, emotional biases such as anticipation of regret and the endowment effect make people opt for sticking. On the other hand, it has also been suggested that the incomplete representation of the different possible courses of action is normally mediated by ignorance about the constraints involved in the elimination of one option. Nevertheless, as argued in this review, the initial partition between three equally likely options instead of two unequal sets of possibilities (contestant's and host's possibilities) seems also to be an important determinant, frequently ignored, for the difficulty in overcoming the equiprobability illusion in the final two-choice scenario.

The relevance of ensuring a correct initial partition of the probability space, combined with understanding that there is a null option within the non-selected partition, is supported by the observation that the best interventions shown to improve Bayesian reasoning in the MHD are the ones promoting the representation of the possibilities of each actor (contestant and host). Furthermore, the dissociation observed between the interventions enhancing optimal choice (repeated practice or increased incentives) and the ones enhancing correct reasoning (explicit partitioning of possibilities) is coherent with current dual process theories of thinking (e.g., Sloman, 1996; Evans, 2010; Kahneman, 2011; Stanovich, 2011) and with dual process models of reward learning (Dayan and Daw, 2008). Whereas changes in preference would be controlled by the autonomous mind (i.e., by means of model-free reward learning mechanisms), explicit reasoning would depend on available cognitive resources and explicit knowledge of the task (similarly to the requirements of model-based reward mechanisms). Accordingly, the present review highlights promising new avenues to help understand behavior and reasoning gaps, and to anticipate the efficacy of new interventions to improve Bayesian reasoning.

Acknowledgments

The work was supported by grant PSI2013-41568-P from MINECO and 2014 SGR-79 from the Catalan Government

References

Ayal, S., and Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgm. Decis. Mak.* 9, 226–242.

Baratgin, J., and Politzer, G. (2010). Updating: a psychologically basic situation of probability revision. *Think. Reason.* 16, 253–287. doi: 10.1080/13546783.2010.519564

Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241–254. doi: 10.1017/S0140525X07001653

Brase, G. L., Fiddick, L., and Harries, C. (2006). Participant recruitment methods and statistical reasoning performance. *Q. J. Exp. Psychol.* 59, 965–976. doi: 10.1080/02724980543000132

Burns, B. D., and Wieth, M. (2004). The collider principle in causal reasoning: why the Monty Hall Dilemma is so hard. *J. Exp. Psychol. Gen.* 133, 434–449. doi: 10.1037/0096-3445.133.3.434

Dayan, P., and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cogn. Affect. Behav. Neurosci.* 8, 429–453. doi: 10.3758/CABN.8.4.429

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: some clarifications. *Think. Reason.* 20, 169–187. doi: 10.1080/13546783.2013.854725

De Neys, W., and Verschueren, N. (2006). Working memory capacity and a notorious brain teaser: the case of the Monty Hall Dilemma. *Exp. Psychol.* 53, 123–131. doi: 10.1027/1618-3169.53.1.123

Evans, J. St. B. T. (2010). *Thinking Twice: Two Minds in One Brain*. Oxford: Oxford University Press.

Evans, J. St. B. T., and Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685

Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. *Cognition* 43, 197–223. doi: 10.1016/0010-0277(92)90012-7

Falk, R., and Lann, A. (2008). The allure of equality: uniformity in probabilistic and statistical judgment. *Cogn. Psychol.* 57, 293–334. doi: 10.1016/j.cogpsych.2008.02.002

Franco-Watkins, A., Derks, P., and Dougherty, M. (2003). Reasoning in the Monty Hall problem: examining choice behavior and probability judgments. *Think. Reason.* 9, 67–90. doi: 10.1080/13546780244000114

Friedman, D. (1998). Monty Hall's three doors: construction and deconstruction of a choice anomaly. *Am. Econ. Rev.* 88, 933–946.

Gigerenzer, G. (1994). "Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa)," in *Subjective Probability*, eds G. Wright and P. Ayton (Oxford: John Wiley & Sons Inc.), 129–161.

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gilovich, T., Medvec, V. H., and Chen, S. (1995). Commission, omission, and dissonance reduction: coping with regret in the "Monty Hall" problem. *Pers. Soc. Psychol. Bull.* 21, 182–190. doi: 10.1177/0146167295212008

Girotto, V., and Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of question form and information structure. *Cognition* 78, 247–276. doi: 10.1016/S0010-0277(00)00133-5

Girotto, V., and Gonzalez, M. (2005). "Probabilistic reasoning and combinatorial analysis," in *The Shape of Reason,* eds V. Girotto and P. N. Johnson-Laird (Hove: Psychology Press).

Glymour, C. N. (2001). *The Mind's Arrow: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: MIT Press.

Granberg, D., and Brown, T. A. (1995). The Monty Hall dilemma. *Pers. Soc. Psychol. Bull.* 21, 711–723. doi: 10.1177/0146167295217006

Granberg, D., and Dorr, N. (1998). Further exploration of two stage decision making in the Monty Hall dilemma. *Am. J. Psychol.* 111, 561–579. doi: 10.2307/1423551

Herbranson, W. T., and Schroeder, J. (2010). Are birds smarter than mathematicians? Pigeons (*columba livia*) perform optimally on a version of the Monty Hall dilemma. *J. Comp. Psychol.* 124, 1–13. doi: 10.1037/a0017703

Herbranson, W. T., and Wang, S. (2014). Testing the limits of optimality: the effect of base rates in the Monty Hall dilemma. *Learn. Behav.* 42, 69–82. doi: 10.3758/s13420-013-0126-6

Herrnstein, R. J. (2000). *The Matching Law: Papers in Psychology and Economics.* Cambridge: Harvard University Press.

Howard, J. N., Lambdin, C. G., and Datteri, D. L. (2007). Let's make a deal: quality and availability of second-stage information as a catalyst for change. *Think. Reason.* 13, 248–272. doi: 10.1080/13546780600848049

Johnson, E. D., and Tubau, E. (2013). Words, numbers, & numeracy: diminishing individual differences in Bayesian reasoning. *Learn. Individ. Differ.* 28, 34–40. doi: 10.1016/j.lindif.2013.09.004

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Sonino-Legrenzi, M., and Caverni, J.-P. (1999). Naive probability: a model theory of extensional reasoning. *Psychol. Rev.* 106, 62–88. doi: 10.1037/0033-295X.106.1.62

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., Knetsch, J., and Thaler, R. H. (1991). Anomalies: the endowment effect, loss aversion and status quo bias. *J. Econ. Perspect.* 5, 193–206. doi: 10.1257/jep.5.1.193

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291. doi: 10.2307/1914185

Klein, E. D., Evans, T. A., Schultz, N. B., and Beran, M. J. (2013). Learning how to "Make a Deal": human (*Homo sapiens*) and monkey (*Macaca mulatta*) performance when repeatedly faced with the Monty Hall Dilemma. *J. Comp. Psychol.* 127, 103–108. doi: 10.1037/a0029057

Koehler, D. J., and James, G. (2010). Probability matching and strategy availability. *Mem. Cognit.* 38, 667–676. doi: 10.3758/MC.38.6.667

Krauss, S., and Wang, X. T. (2003). The psychology of the Monty Hall problem: discovering psychological mechanisms for solving a tenacious brain teaser. *J. Exp. Psychol. Gen.* 132, 3–22. doi: 10.1037/0096-3445.132.1.3

Krynski, T. R., and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *J. Exp. Psychol. Gen.* 136, 430–450. doi: 10.1037/0096-3445.136.3.430

Landman, J. (1988). Regret and elation following action and inaction: affective responses to positive versus negative outcomes. *Pers. Soc. Psychol. Bull.* 13, 524–536. doi: 10.1177/0146167287134009

Langer, E. (1975). The illusion of control. *J. Pers. Soc. Psychol.* 32, 311–328. doi: 10.1037/0022-3514.32.2.311

Legrenzi, P., Girotto, V., and Johnson-Laird, P. N. (1993). Focusing in reasoning and decision making. *Cognition* 48, 37–66. doi: 10.1016/0010-0277(93)90035-T

Leotti, L. A., Iyengar, S. S., and Ochsner, K. N. (2010). Born to choose: the origins and value of the need for control. *Trends Cogn. Sci.* 14, 457–463. doi: 10.1016/j.tics.2010.08.001

Lichtenstein, S., and Slovic, P. (1971). Reversal of preferences between bids and choices in gambling decisions. *J. Exp. Psychol.* 89, 46–55. doi: 10.1037/h0031207

Mazur, J. E., and Kahlbaugh, P. E. (2012). Choice behavior of pigeons (*Columba livia*), college students, and preschool children (*Homo sapiens*) in the Mony Hall Dilemma. *J. Comp. Psychol.* 126, 407–420. doi: 10.1037/a0028273

McNair, S., and Feeney, A. (2014). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychon. Bull. Rev.* 22, 258–264. doi: 10.3758/s13423-014-0645-y

Morone, A., and Fiore, A. (2008). "Monty Hall's three doors for dummies," in *Advances in Decision Making Under Risk and Uncertainty Theory and Decision Library,* Vol. 42, eds M. Abdellaoui and J. D. Hey (Berlin: Springer), 151–162.

Nowak, M., and Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's dilemma game. *Nature* 364, 56–58. doi: 10.1038/364056a0

Palacios-Huerta, I. (2003). Learning to open Monty Hall's doors. *Exp. Econ.* 6, 235–251. doi: 10.1023/A:1026209001464

Petrocelli, J. V., and Harris, A. K. (2011). Learning inhibition in the Monty Hall Problem: the role of dysfunctional counterfactual prescriptions. *Pers. Soc. Psychol. Bull.* 37, 1297–1311. doi: 10.1177/0146167211410245

Piattelli-Palmarini, M. (1994). *Inevitable Illusions: How Mistakes of Reason Rule Our Minds*. New York, NY: Wiley.

Saenen, L., Van Dooren, W., and Onghena, P. (2014). A randomised Monty Hall experiment: the positive effect of conditional frequency feedback. *Think. Reason.* 1–17. doi: 10.1080/13546783.2014.918562 [Epub ahead of print].

Shimojo, S., and Ichikawa, S. (1989). Intuitive reasoning about probability: theoretical and experimental analysis of the "problem of three prisoners". *Cognition* 32, 1–24. doi: 10.1016/0010-0277(89)90012-7

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychol. Bull.* 119, 3–22. doi: 10.1037/0033-2909.119.1.3

Stanovich, K. E. (2009). "Distinguishing the reflective, algorithmic, and autonomous minds: is it time for a tri-process theory?" in *In Two Minds: Dual Processes and Beyond,* eds J. Evans and K. Frankish (Oxford: Oxford University Press), 55–88. doi: 10.1093/acprof:oso/9780199230167.003.0003

Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. New York, NY: Oxford University Press.

Stibel, J. M., Dror, I. E., and Ben-Zeev, T. (2009). Dissociating choice and judgment in decision making: the collapsing choice theory. *Theor. Decis.* 22, 149–179. doi: 10.1007/s11238-007-9094-7

Stocco, A., and Fum, D. (2008). Implicit emotional biases in decision making: the case of the Iowa gambling task. *Brain Cogn.* 66, 253–259. doi: 10.1016/j.bandc.2007.09.002

Thompson, V. A. (2009). "Dual process theories: a metacognitive perspective," in *Two Minds: Dual Processes and Beyond*, eds J. Evans and K. Frankish (Oxford: Oxford University Press).

Tor, A., and Bazerman, M. H. (2003). Focusing failures in competitive environments: explaining decision errors in the Monty Hall game, the acquiring a company problem, and multiparty ultimatums. *J. Behav. Decis. Making* 16, 353–374. doi: 10.1002/bdm.451

Tubau, E. (2008). Enhancing probabilistic reasoning: the role of causal graphs, statistical format and numerical skills. *Learn. Individ. Differ.* 18, 187–196. doi: 10.1016/j.lindif.2007.08.006

Tubau, E., and Alonso, D. (2003). Overcoming illusory inferences in a probabilistic counterintuitive problem: the role of explicit representations. *Mem. Cogn.* 31, 596–607. doi: 10.3758/BF03196100

Tversky, A., and Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychol. Rev.* 101, 547–567. doi: 10.1037/0033-295X.101.4.547

van der Pligt, J. (1998). Perceived risk and vulnerability as predictors of precautionary health behaviour. *Br. J. Health Psychol.* 3, 1–14. doi: 10.1111/j.2044-8287.1998.tb00551.x

West, R. F., and Stanovich, K. E. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Mem. Cognit.* 31, 243–251. doi: 10.3758/BF03194383

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in PSYCHOLOGY

OPINION ARTICLE
published: 29 October 2014
doi: 10.3389/fpsyg.2014.01232

Visual representation of rational belief revision: another look at the Sleeping Beauty problem

David R. Mandel*

Socio-Cognitive Systems Section, Defence Research and Development Canada, Toronto Research Centre, Department of Psychology, York University, Toronto, ON, Canada
**Correspondence: drmandel66@gmail.com*

Edited by:
Gorka Navarrete, Universidad Diego Portales, Chile

Reviewed by:
David E. Over, Durham University, UK
Jean Baratgin, Université Paris 8, France

Keywords: Bayesian reasoning, belief revision, visual representation, rationality, Sleeping Beauty problem

The coherence of probability judgments is influenced in predictable ways by people's internal representations of problems, which may be altered by the manner in which propositions are stated or "framed" (Mandel, 2008). Likewise, several studies find that probabilistic reasoning and judgment can be improved by externally representing statistical information visually (for a review, see Garcia-Retamero and Cokely, 2013). Visual representation is thought to facilitate performance by externalizing the set-subset relations among observational data. Although some studies have examined whether visual representations can improve Bayesian reasoning, they have tended to focus on the use of natural sampling trees (Sedlmeier and Gigerenzer, 2001), Euler circles (Sloman et al., 2003), or other means of representing set-subset relations.

However, visualization can aid reasoning and judgment even when problems do not involve natural or normalized frequency representations. Take the "Ann problem" adapted by Over (2007b):

> Jack is looking at Ann but Ann is looking at George. Jack is a cheater but George is not. Is a cheater looking at a non-cheater?
> (A) Yes (B) No (C) Cannot tell

In a variant of the problem, Toplak and Stanovich (2002) found that most people say they cannot tell, although the correct answer is yes. Wrong answers are common because most people do not consider the implications of the fact that Ann is either a cheater or she is not. As Over (2007b) notes, the logic of the excluded middle—namely, that all propositions of the form "*x* or not-*x*" are logically true—is often neglected.

Instead people seem to be guided by their sense of uncertainty about both of the dyadic relations in the problem, remaining unaware that their uncertainty should not preclude a more definite conclusion. As Over (2007a,b) suggests, logic trees, which represent possibilities on branches, can provide a useful visualization tool for overcoming such psychological barriers. If one were to draw out the two possibilities in the Ann problem—one in which cheater Jack looks at non-cheater Ann and the other in which cheater Ann looks at non-cheater George—the correct answer is evident. If you draw a logic tree showing the two possibilities (Ann as a cheater or as a non-cheater) and the "looking relations" that are entailed in each, it becomes evident that no matter what Ann is, a cheater will always look at a non-cheater. Who the cheater is and who the non-cheater is will differ depending on whether Ann is a cheater or not, but those details are irrelevant to the question. The logic tree also shows that it is impossible for a non-cheater to look at a cheater. However, in that case, one must attend to what is omitted from the set of possible worlds.

THE SLEEPING BEAUTY PROBLEM

In the remainder of this paper, I explore the value of logic trees in representing alternative arguments by experts about normative belief updating. I focus on the Sleeping Beauty problem introduced by Elga (2000) and discussed shortly thereafter by Lewis (2001). My aim is twofold: First, I want to show how these authors' arguments may be represented and how the representations may be compared. Second, I want to propose a resolution of the disagreement over the problem that I believe is novel.

This is Lewis's description of the problem:

> Researchers at Experimental Philosophy Laboratory have decided to carry out the following experiment. First they will tell Sleeping Beauty [SB] all that I am about to tell you in this paragraph, and they will see to it that she fully believes all she is told. Then on Sunday evening they will put her to sleep. On Monday they will awaken her briefly. At first they will not tell her what day it is, but later they will tell her that it is Monday. Then they will subject her to memory erasure. Perhaps they will again awaken her briefly on Tuesday. Whether they do will depend on the toss of a fair coin: if heads they will awaken her only on Monday, if tails they will awaken her on Tuesday as well. On Wednesday the experiment will be over and she will be allowed to wake up. The three possible brief awakenings during the experiment will be indistinguishable: she will have the same total evidence at her Monday awakening whatever the result of the coin toss may be, and if she is awakened on Tuesday the memory erasure on Monday will make sure that her total evidence at the Tuesday awakening is exactly the same as at the Monday awakening. However, she will be able, and she will be taught how, to distinguish her brief awakenings during the experiment

> from her Wednesday awakening after the experiment is over, and indeed from all other actual awakenings there have ever been, or ever will be.

Furthermore, assume that SB is a paragon of rationality and let us also assume for the sake of concreteness that the coin is tossed on Sunday night after SB is put to sleep. What subjective probability should she assign to heads (H) upon her awakening on Monday, and then again after she is told that it is Monday?

Elga and Lewis agree that SB will be in one of three states:

> H_1: Heads and it is Monday
> T_1: Tails and it is Monday
> T_2: Tails and it is Tuesday.

Elga starts out by imagining that SB knows that the coin lands on tails. Since T_1 and T_2 would be indistinguishable to SB, he argues that she should assign each the same probability: $P(T_1) = P(T_2) = 1/2$. Next, Elga imagines that SB knows it is Monday, arguing that SB should assign equal probability to H_1 and T_1 given the fact that the coin is fair. Thus, $P(H_1) = P(T_1) = P(T_2)$. Since these probabilities must sum to 1, each must equal 1/3. Therefore, Elga proposes that, on waking in an asynchronous state, SB should assign a 1/3 probability to heads, and that she should revise this probability to 1/2 after learning it is Monday.

Lewis disagrees. He starts out with the principle that the subjective probability of a future chance event should be equal to the known chances (Mellor, 1971; Lewis, 1980). Since the coin is fair, the known chances indicate $P(H) = P(T) = 1/2$. Lewis argues that on awakening SB has not learned anything new that would warrant belief revision. She has no new knowledge of her location. Like Elga, Lewis accepts that SB should regard $P(T_1) = P(T_2)$. Given $P(T) = P(T_1 \vee T_2) = 1/2$, and the disjunctive possibilities are equiprobable, $P(T_1) = P(T_2) = 1/4$.

Elga and Lewis agree that, upon learning it is Monday, SB should increase her subjective probability of heads by 1/6 after conditionalizing on the remaining possibilities. For Elga, $P(H|H_1 \vee T_1) = (1/3)/(2/3) = 1/2$. For Lewis, $P(H \,|\, H_1 \vee T_1) = (1/2)/(3/4) = 2/3$. Interestingly, Lewis does not apply his imaging rule for belief updating (Lewis, 1976) here, even though it arguably applies (Cozic, 2011; see also Baratgin, 2009).

The SB problem continues to prompt philosophical debate (e.g., Dorr, 2002; Horgan, 2004; Weintraub, 2004; Rosenthal, 2009; Baratgin and Walliser, 2010). In my own thinking about it, I have found it useful to externally visualize the alternative arguments using enhanced logic trees that also encode operations (e.g., normalization) or relation types (e.g., necessity). **Figure 1** shows possible logic trees for Elga's "thirder" and Lewis's "halfer" positions. It reveals that the locus of disagreement is in the apportioning of probability to T_1 and T_2.

In Elga's analysis, these two centered possibilities each have a subjective probability of 1/2 since the coin toss outcome T, all agree, equals 1/2 and the Monday and Tuesday awakenings necessarily follow. Since H_1 also equals 1/2, the probabilities must be normalized to constrain their sum to 1. This leads to each centered possibility having a probability of 1/3.

In Lewis's analysis, the same two centered possibilities, T_1 and T_2, each have a subjective probability of 1/4 because Lewis applies a principle of indifference to them. Given that the three centered possibilities are additive, normalization is not required and H_1 remains 1/2.

The visualizations reveal something about the relative strength of the two positions, which I believe favors 1/3 as an *answer* to the first question. I won't say they favor Elga's *arguments* over Lewis's. That would be reading in too much and let me come back to that. It seems evident that the strength of the Elgan tree over the Lewisian tree is that the former encodes necessity relations on the centered branches that follow from the possible world in which T transpired on Sunday night, whereas the latter encodes SB's uncertainties. We have already seen what relying on our uncertainties rather than on what must follow can do in the Ann problem. I suspect the lesson may be repeated here but for better reasons. Lewis keeps $P(H_1)$ fixed at 1/2 because he believes that, given no change in relevant information, there should not be a change in subjective probability. Since all agree that $P(H) = 1/2$, and since nothing about location is learned upon SB's awakening, there *is* a principled reason for not changing the probabilities. As Lewis notes, he realizes the appeal of Elga's argument, but it is precisely because he finds his own more principled that he sticks to it. There is something to be said about following logic even if it does not lead to intuitive conclusions, and that appears to be what Lewis has done.

While Lewis is correctly principled, both he and Elga mistake what SB's subjective probability on Sunday ought to refer to. Both attribute a subjective probability of 1/2 to SB on Sunday night before she is put to sleep. But what exactly does this probability refer to? Elga and Lewis focus on $P(H)$, and I believe that is the problem. One should consider what probability SB would assign on Sunday to H knowing what she knows about the waking rules of the experiment, and imagining she has just awoken in an asynchronous state in the experiment. Let us call this $P^*(H_1)$, where the asterisk denotes the counterfactual status of the hypothesis. $P^*(H_1)$ is the probability of the Stalnaker-type conditional (Stalnaker, 1968) specified in the query, "What is the probability that if you, SB, were to have an asynchronous awakening, then the coin would have come up heads?" We might expand this query, which utilizes a wide-scope probability operator (Over et al., 2013), as follows: "What is the probability that if you, SB, were to have an asynchronous awakening, which in fact you and I know you are not having at the moment, and if you knew all that you know now about the rules of the experiment, then the coin would have come up heads?" In this case, the probability she should assign to $P^*(H_1)$ equals 1/3, precisely because $P(H) = P(T) = 1/2$, P(Monday awakening) = 1, and P(Tuesday awakening) = 1/2. Because an asynchronous awakening, A, must either be a Monday awakening or a Tuesday awakening, $P(A = \text{Monday}) = 2/3$. $P^*(H_1) = P(A = \text{Monday})P(H) = (2/3)(1/2) = 1/3$.

That, on Monday, $P(H_1)$ should also equal 1/3 reflects adherence to the dynamic coherence criterion or Bayesian conditioning principle, which states that a probability assessed conditionally on a suppositional event x should not differ from the probability assessed conditionally

on the actual event x (Baratgin and Politzer, 2006). In the Sleeping Beauty problem, A may be supposed, contrary to fact, on Sunday night and A will be actualized on Monday, and possibly on Tuesday too.

The mislabeling of the event that SB is to consider on Sunday night leads Elga to accept belief revision in the absence of new relevant information. He arrives at a correct answer but forfeits a principle he should have defended. Lewis defends that principle, but ends up with an incorrect estimate because of the initial labeling error. Elgan thirders are therefore right about 1/3 and Lewisian halfers are right to stick to their principles.

Both the Ann and Sleeping Beauty problems illustrate the value of visual representations in reasoning through problems that require people to state their degree of belief in a given proposition. In neither case is the problem's solution clarified by externalizing a natural frequency representation of the problem. Frequency trees and other nested-set-revealing visualizations may facilitate Bayesian reasoning, but so can other forms of visualization, such as (enhanced) logic trees.

The Sleeping Beauty problem also highlights the limits of visualization since nothing in the visualizations offered clarifies the labeling error that I believe lies at the heart of the disagreement; namely, that the proposition being assessed changes from Time 1 (Sunday night) to Time 2 (Monday's asynchronous awakening). Put differently, the visualizations shown in **Figure 1** do not represent queries, and it is at the level of query formulation where I believe the controversy first arose. Note too that while the trees in **Figure 1** respectively represent Elga's and Lewis's stances on the Sleeping Beauty problem, they do not inherently resolve which stance is more appropriate. At best, they might help other reasoners reach a conclusion by showing in representational terms where disagreement seems to lie.

If my account is correct, it raises the question why $P^*(H_1)$ could be mistaken for $P(H)$ by such sharp minds. That it would—namely, that Sunday's apples would be compared with Monday's oranges—is both surprising and a continuing source of my own skepticism in its correctness. Yet, it seems uncontroversial that (a) Elga, Lewis and indeed most commentators on the problem focus their attention on $P(H)$ when considering SB's Sunday assessment and (b) that this is not well paired with the assessments made upon awakening. To be explicit, the reason it is not well paired is that on Monday, SB must take into account the rules of the experiment, which she perfectly remembers, yet on Sunday she must disregard that knowledge, which is equally at her disposal, in giving her simple credence for heads. Given she is a paragon of rationality, I cannot help but think that she would object to such inconsistency.

ACKNOWLEDGMENTS

I thank Jean Baratgin and David Over for helpful comments on an earlier draft of this paper and for engaging me in useful

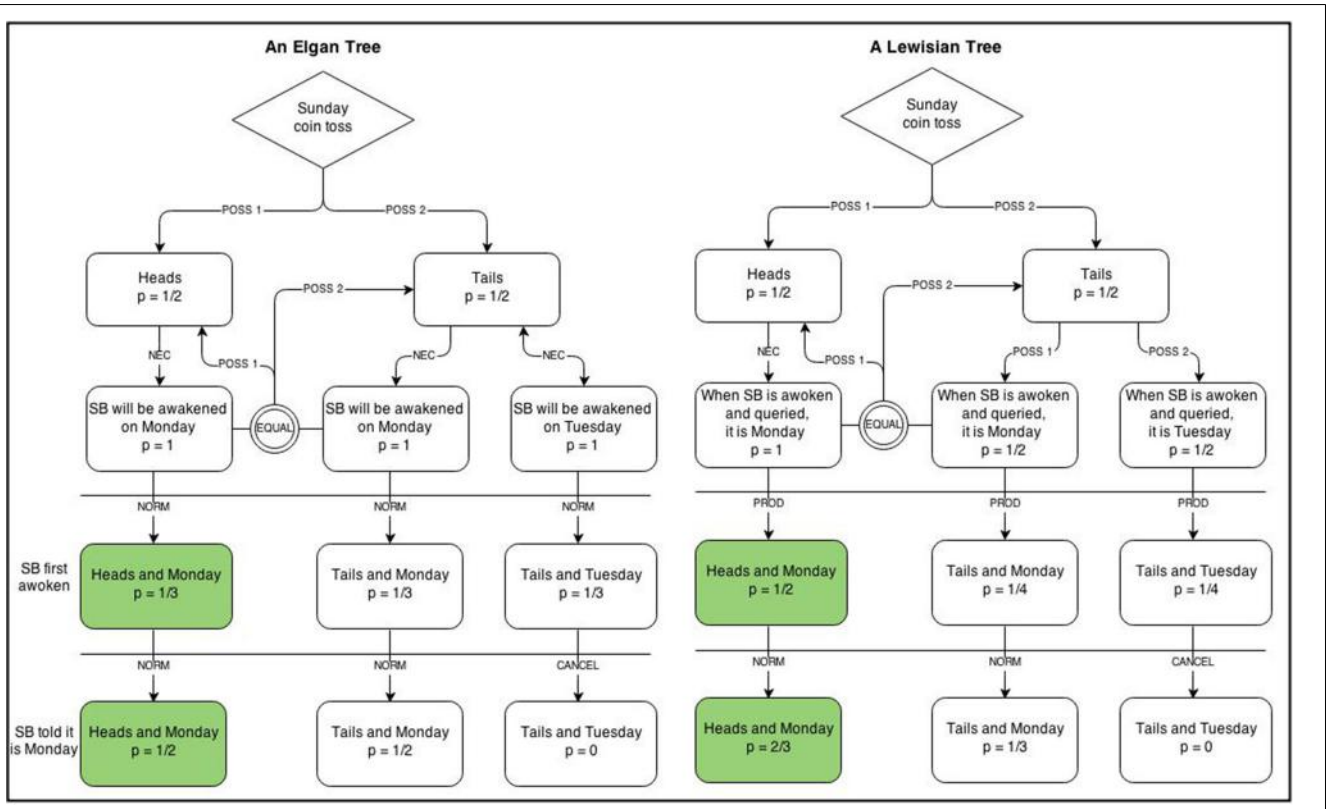

FIGURE 1 | Enhanced logic trees for Elga's (2000) and Lewis's (2001) positions on the Sleeping Beauty problem. POSS, possibility; NEC, necessity; NORM, normalize; PROD, product.

discussions about the Sleeping Beauty problem and belief revision.

REFERENCES

Baratgin, J. (2009). Updating our beliefs about inconsistency: the Monty Hall case. *Math. Soc. Sci.* 57, 67–95. doi: 10.1016/j.mathsocsci.2008.08.006

Baratgin, J., and Politzer, G. (2006). Is the mind Bayesian? The case for agnosticism. *Mind Soc.* 5, 1–38. doi: 10.1007/s11299-006-0007-1

Baratgin, J., and Walliser, B. (2010). Sleeping Beauty and the absent-minded driver. *Theory Decis.* 69, 489–496. doi: 10.1007/s11238-010-9215-6

Cozic, M. (2011). Imaging and Sleeping Beauty: the case for double-halfers. *Int. J. Approx. Reason.* 52, 147–153. doi: 10.1016/j.ijar.2009.06.010

Dorr, C. (2002). Sleeping Beauty: in defence of Elga. *Analysis* 62, 292–296. doi: 10.1093/analys/62.4.292

Elga, A. (2000). Self-locating belief and the Sleeping Beauty problem. *Analysis* 60, 143–147. doi: 10.1093/analys/60.2.143

Garcia-Retamero, R., and Cokely, E. T. (2013). Communicating health risks with visual aids. *Curr. Dir. Psychol. Sci.* 22, 392–399. doi: 10.1177/0963721413491570

Horgan, T. (2004). Sleeping Beauty awakened: new odds at the dawn of the new day. *Analysis* 64, 10–21. doi: 10.1093/analys/64.1.10

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philos. Rev.* 85, 297–315. doi: 10.2307/2184045

Lewis, D. (1980). "A subjectivist's guide to objective chance," in *Studies in Inductive Logic and Probability*, Vol. 2, ed R. C. Jeffrey (Oxford: Oxford University Press), 263–293.

Lewis, D. (2001). Sleeping Beauty: reply to Elga. *Analysis* 61, 171–176. doi: 10.1093/analys/61.3.171

Mandel, D. R. (2008). Violations of coherence in subjective probability: a representational and assessment processes account. *Cognition* 106, 130–156. doi: 10.1016/j.cognition.2007.01.001

Mellor, D. H. (1971). *The Matter of Chance*. Cambridge: Cambridge University Press.

Over, D. E. (2007a). "Content-independent conditional inference," in *Integrating the Mind: Domain General Versus Domain Specific Processes in Higher Cognition*, ed M. J. Roberts (New York, NY: Psychology Press), 83–103.

Over, D. E. (2007b). The logic of natural sampling. *Behav. Brain Sci.* 30:277. doi: 10.1017/S0140525X07001859

Over, D. E., Dougen, I., and Verbrugge, S. (2013). Scope ambiguities and conditionals. *Think. Reason.* 19, 284–307. doi: 10.1080/13546783.2013.810172

Rosenthal, J. S. (2009). A mathematical analysis of the Sleeping Beauty problem. *Math. Intell.* 31, 32–37. doi: 10.1007/s00283-009-9060-z

Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* 130, 380–400. doi: 10.1037/0096-3445.130.3.380

Sloman, S. A., Over, D. E., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Process.* 91, 296–309. doi: 10.1016/S0749-5978(03)00021-9

Stalnaker, R. (1968). "A theory of conditionals," in *Studies in Logical Theory*, ed N. Rescher (Oxford, UK: Blackwell), 98–112.

Toplak, M. E., and Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning. Searching for a generalizable critical reasoning skill. *J. Educ. Psychol.* 94, 197–209. doi: 10.1037/0022-0663.94.1.197

Weintraub, R. (2004). Sleeping Beauty: a simple solution. *Analysis* 64, 8–10. doi: 10.1093/analys/64.1.8

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 16 September 2014; accepted: 10 October 2014; published online: 29 October 2014.

Citation: Mandel DR (2014) Visual representation of rational belief revision: another look at the Sleeping Beauty problem. Front. Psychol. ***5****:1232. doi: 10.3389/fpsyg.2014.01232*

This article was submitted to Cognition, a section of the journal Frontiers in Psychology.

ADVANTAGES OF PUBLISHING IN FRONTIERS

FAST PUBLICATION
Average 90 days from submission to publication

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

RESEARCH NETWORK
Our network increases readership for your article

OPEN ACCESS
Articles are free to read, for greatest visibility

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

GLOBAL SPREAD
Six million monthly page views worldwide

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

IMPACT METRICS
Advanced metrics track your article's impact

SUPPORT
By our Swiss-based editorial team

frontiers

EPFL Innovation Park · Building I · 1015 Lausanne · Switzerland
T +41 21 510 17 00 · info@frontiersin.org · frontiersin.org