JUDGMENT AND DECISION MAKING UNDER UNCERTAINTY: DESCRIPTIVE, NORMATIVE, AND PRESCRIPTIVE PERSPECTIVES

EDITED BY: David R. Mandel, Gorka Navarrete, Nathan Dieckmann and Jonathan D. Nelson

PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88963-034-9
DOI 10.3389/978-2-88963-034-9

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

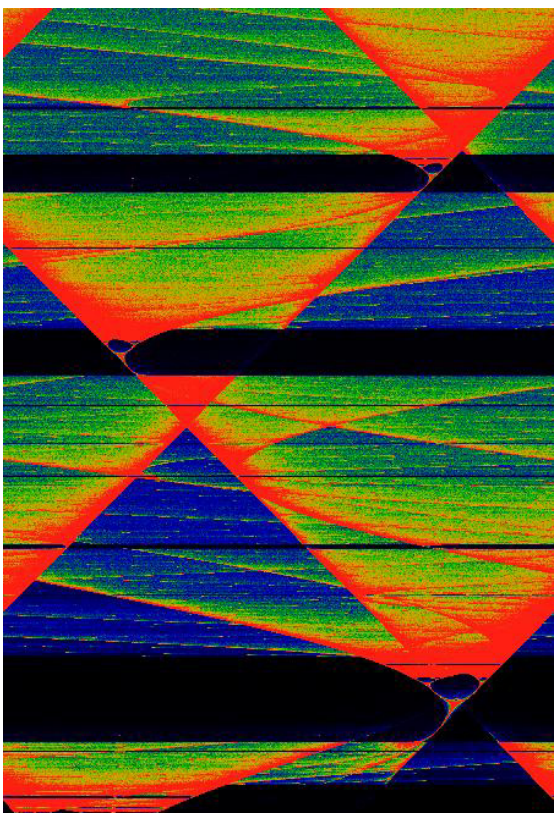
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - 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

JUDGMENT AND DECISION MAKING UNDER UNCERTAINTY: DESCRIPTIVE, NORMATIVE, AND PRESCRIPTIVE PERSPECTIVES

Topic Editors:
David R. Mandel, Defence Research and Development Canada, Canada; York University, Canada
Gorka Navarrete, Universidad Adolfo Ibáñez, Chile
Nathan Dieckmann, Oregon Health & Science University, United States; Decision Research, United States
Jonathan D. Nelson, University of Surrey, United Kingdom; Max Planck Institute for Human Development, Germany



Circle Map Bifurcation

Citation: Mandel, D. R., Navarrete, G., Dieckmann, N., Nelson, J. D., eds. (2019). Judgment and Decision Making Under Uncertainty: Descriptive, Normative, and Prescriptive Perspectives. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88963-034-9

Table of Contents

frontiers in Psychology

EDITORIAL
published: 02 July 2019
doi: 10.3389/fpsyg.2019.01506

Editorial: Judgment and Decision Making Under Uncertainty: Descriptive, Normative, and Prescriptive Perspectives

David R. Mandel[1,2]*, *Gorka Navarrete*[3], *Nathan Dieckmann*[4,5] and *Jonathan Nelson*[6,7]

[1] Intelligence, Influence and Collaboration Section, Defence Research and Development Canada, Toronto, ON, Canada, [2] Department of Psychology, York University, Toronto, ON, Canada, [3] Center for Social and Cognitive Neuroscience, School of Psychology, Universidad Adolfo Ibáñez, Santiago, Chile, [4] School of Nursing, Oregon Health & Science University, Portland, OR, United States, [5] Decision Research, Eugene, OR, United States, [6] School of Psychology, University of Surrey, Guildford, United Kingdom, [7] iSearch Group, Max Planck Institute for Human Development, Berlin, Germany

Keywords: judgment, decision-making, uncertainty, cognition, psychology

Editorial on the Research Topic

Judgment and Decision Making Under Uncertainty: Descriptive, Normative, and Prescriptive Perspectives

Judgment and Decision Making Under Uncertainty: Descriptive, Normative, and Prescriptive Perspectives was motivated by our interest in better understanding why people judge and decide as they do (descriptive perspective), how they ideally ought to judge and decide (normative perspective), and how their judgment and decision-making processes might be improved in practice (prescriptive perspective). We sought papers that addressed some aspect of judgment and decision making from one or more of these three theoretical perspectives. We further sought contributions that examined judgment and decision making under conditions of uncertainty, which we intentionally left loosely defined. Our focus on uncertainty reflects the fact that the vast majority of decisions people make in life are not made under conditions of complete certainty, and the uncertainties may be more or less well-defined. Indeed, different components of a single judgment or decision may have multiple uncertainties associated with it, some of which may be fuzzier than others. Following our call for papers, we received 32 submissions, 17 of which were accepted. The latter set comprises this book. There are 11 original research articles, 2 hypothesis and theory articles, 2 perspectives, and 1 book review and systematic review each.

This book, the culmination of a *Frontiers in Psychology* Cognition section Research Topic, is closely related to an earlier Research Topic and book entitled *Improving Bayesian Reasoning: What Works and Why* that two of us edited (Navarrete and Mandel, 2016). The current book shows strong continuity with its conceptual cousin. Several papers address aspects of Bayesian judgment or reasoning. In "Why can only 24% solve Bayesian reasoning problems in natural frequencies: Frequency phobia in spite of probability blindness," Weber et al. find that, despite the benefit to accuracy conferred by representing statistical information in natural frequencies, many participants translate natural frequencies back into probabilities. This appears to be an important factor in explaining the low rates of accurate judgment. In "How to improve performance in Bayesian inference tasks: A comparison of five visualizations," Böcherer-Linder and Eichler investigate the effectiveness of three graphical properties of visualizations: area-proportionality, use of discrete and countable statistical entities, and graphical transparency of the nested-sets structure. They find that the primary factor contributing to performance in Bayesian reasoning

OPEN ACCESS

Edited and reviewed by:
Bernhard Hommel,
Leiden University, Netherlands

***Correspondence:**
David R. Mandel
drmandel66@gmail.com

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

Received: *07 June 2019*
Accepted: *14 June 2019*
Published: *02 July 2019*

Citation:
Mandel DR, Navarrete G, Dieckmann N and Nelson J (2019) Editorial: Judgment and Decision Making Under Uncertainty: Descriptive, Normative, and Prescriptive Perspectives. Front. Psychol. 10:1506. doi: 10.3389/fpsyg.2019.01506

problems was graphically representing the nested-set structure of the problem in a transparent manner, followed secondarily by representing discrete objects. In "What eye-tracking can tell us on statistical reasoning—An empirical study on tree diagrams and 2 × 2 tables," Bruckmaier et al. use eye tracking to shed light on the reasons for errors in probabilistic judgment. They show that different reasoning processes can account for errors that look similar behaviorally. Conversely, errors that look different may stem from common reasoning processes. In "Bayesian revision vs. information distortion," Russo explains how a normative requirement of Bayesian reasoning—namely, that likelihoods should be independent of the prior probability—is routinely violated in all but the most contrived judgment problems where such violations are designed to be impossible. The violations, Russo argues, occur because people strive for coherence and therefore seek to bring new evidence in line with their prior beliefs. Evidently, the pursuit of coherence can at times signal its downfall. Finally, in "Metacognitive myopia in hidden-profile tasks: the failure to control for repetition biases," Fiedler et al. address an issue that is conceptually related to updating processes when confronting correlated evidence. They find that mere repetition of information over time (which can be thought of as a form of correlated evidence) can undermine the optimal use of information that is distributed across members of a collective. As they aptly point out, given the vast opportunities for information repetition to trigger such biases, it is vital that metacognitive monitoring takes place, and yet their results indicate that people have a difficult time doing so.

A second set of papers tackles uncertainty from several fresh vantage points. In "The psychology of uncertainty and three-valued tables," Baratgin et al. examine people's three-valued (i.e., certainly true, certainly false, or neither) truth tables for several natural language connectives. Comparing multiple three-valued logics, they find that de Finetti's (1936/1995) three-valued system provides the best descriptive model. Their work on the de Finettian "Level 1," in which uncertainty is distinguished from certain states, represents a long neglected bridge between Level 0 (binary logic) and Level 2 (studies of probability judgment in which uncertainty is quantified). In "Imprecise uncertain reasoning: a distributional approach," Kleiter develops an approach to using mental probability logic in concert with beta distributions, copulas, vines, and stochastic simulation to model imprecise and uncertain reasoning. A key finding from his analysis of several classic judgment problems is that the probabilities inferred from different logical inference forms can be so close as to make their distinction impossible in psychological research, a result that has striking implications for the interpretation of evidence in judgment research. In "The role of type and source of uncertainty on the processing of climate models projections," Benjamin and Budescu examine how people's interpretations of climate change forecasts from multiple experts are influenced by two sources of uncertainty: imprecision (i.e., the width of the confidence interval around a single estimate) and conflict (the extent to which experts disagree). They find that participants were more averse to conflict and reacted more positively to communications that reflect imprecision. Their results show that people's perceptions of competing climate change forecasts are affected by a complex interaction between sources of uncertainty and task characteristics. This set of papers is nicely rounded out by Mousavi's book review of Machina and Viscusi's *Handbook of the Economics of Risk and Uncertainty*.

A third set of papers addresses topics in decision-making under uncertainty. In "Meta-analytic evidence for a reversal learning effect on the Iowa Gambling Task in older adults," Pasion et al. report a systematic review of studies examining older-adult decision-making on the Iowa Gambling Task. They find evidence of a significant reversal learning effect across blocks of the task, which suggests that older adults show adaptive decision making as they gain experience with the outcomes. In "Cognitive style and frame susceptibility in decision making," Mandel and Kapler examine the predictive effect of several cognitive style and performance measures on frame susceptibility or "going with the frame." They do not find such factors to be predictive of frame susceptibility and they question the theoretical claim that individuals who are prone to a less deliberate, or more intuitive, thinking style are more susceptible to framing effects. In "Too worried to judge: On the role of perceived severity in medical decision-making," Colomé et al. examine content effects on recommendations for medical treatments. They find that worry affects recommendations only in the higher severity context (cancer), whereas consideration of disease likelihood given a positive test result played a greater role in the lower severity context (hypertension). In "The reciprocal relationships between escalation, anger, and confidence in investment decisions over time," Jackson et al. show in an escalation of commitment task, where money had to be invested in different rounds in a never-ending project, people tend to escalate through all rounds. However, as they do, their confidence decreases and anger increases, thus shedding light on the experiential side of this well-documented phenomenon. In "Does fear increase search effort in more numerate people? An experimental study investigating information acquisition in a decision from experience task," Traczyk et al. examine the role of numeracy and emotion of fear on search policy and choice in a decision from experience task. Both numeracy and fear were related to increased information sampling, although the effect of fear was restricted to a more numerate subsample. Their results shed light on the interaction between numeracy and integral emotion in decisions from experience.

Last but not least, three papers draw on decision science to shed light on professional practices in forensics and national security intelligence. In "Decisional dimensions in expert witness testimony—a structural analysis," Biedermann and Kotsoglou integrate decision theory with current practices in forensic science for the use of expert witness testimony. The authors review current theoretical understanding of the expert witness testimony process and then discuss a decision-theoretic framework including real-world examples. In "Better together: reliable application of the post-9/11 and post-Iraq US intelligence tradecraft standards requires collective analysis," Marcoci et al. turn their attention to the US intelligence community's analytic tradecraft standards by asking whether

raters can interpret the standards reliably as they pertain to intelligence products. Overall, the reliability of single raters was poor. Having important prescriptive implications for quality control within the intelligence community, Marcoci et al. find that a group of three or more raters is needed to provide reliable assessments of the quality of intelligence products. Finally, in "Correcting judgment correctives in national security intelligence," Mandel and Tetlock argue that the intelligence community's prescriptions for improving analysts' intelligence assessments—namely, their judgments under uncertainty—could be substantially improved by scientifically testing the effectiveness of proposed methods; something rarely done. Drawing on decision science, Mandel and Tetlock argue that current methods might not only fail to improve analysts' judgments, they may in fact be making intelligence assessments less reliable, coherent and accurate.

We hope the reader will find this book informative, thought provoking, and of practical and theoretical value.

AUTHOR CONTRIBUTIONS

DM wrote the editorial. GN, ND, and JN provided input and constructive feedback.

ACKNOWLEDGMENTS

Support for the preparation of this book was provided by Department of National Defence projects 05da, 05fa and Canadian Safety and Security Program project CSSP-2016-TI-2224 to DM and by a grant from Comisión Nacional de Investigación Científica y Tecnológica (CONICYT/FONDECYT Regular 1171035) to GN.

REFERENCES

de Finetti, B. (1936/1995). La logique de la probabilité. Actes du congrès international de philosophie scientifique. Sorbonne, 1935. IV: induction et probabilité, 31-39. Paris: Hermann. English translation (1995): the logic of probability. *Philos. Stud.* 77, 181–190.

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

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: 12 October 2018
doi: 10.3389/fpsyg.2018.01833

Why Can Only 24% Solve Bayesian Reasoning Problems in Natural Frequencies: Frequency Phobia in Spite of Probability Blindness

Patrick Weber, Karin Binder and Stefan Krauss*

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

For more than 20 years, research has proven the beneficial effect of natural frequencies when it comes to solving Bayesian reasoning tasks (Gigerenzer and Hoffrage, 1995). In a recent meta-analysis, McDowell and Jacobs (2017) showed that presenting a task in natural frequency format increases performance rates to 24% compared to only 4% when the same task is presented in probability format. Nevertheless, on average three quarters of participants in their meta-analysis failed to obtain the correct solution for such a task in frequency format. In this paper, we present an empirical study on what participants typically do wrong when confronted with natural frequencies. We found that many of them did not actually *use* natural frequencies for their calculations, but translated them back into complicated probabilities instead. This switch from the intuitive *presentation format* to a less intuitive *calculation format* will be discussed within the framework of psychological theories (e.g., the Einstellung effect).

Keywords: Bayesian reasoning, natural frequencies, probabilities, einstellung, tree diagram

OPEN ACCESS

Edited by:
Gorka Navarrete,
Adolfo Ibáñez University, Chile

Reviewed by:
Laura Felicia Martignon,
Ludwigsburg University, Germany
Luana Micallef,
University of Copenhagen, Denmark

****Correspondence:***
Patrick Weber
patrick.weber@ur.de

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

Received: *16 March 2018*
Accepted: *07 September 2018*
Published: *12 October 2018*

Citation:
Weber P, Binder K and Krauss S (2018) Why Can Only 24% Solve Bayesian Reasoning Problems in Natural Frequencies: Frequency Phobia in Spite of Probability Blindness. Front. Psychol. 9:1833. doi: 10.3389/fpsyg.2018.01833

INTRODUCTION

Many professionals, such as medical doctors and judges in court, are expected to make momentous decisions based on statistical information. Often, Bayesian inferences are required, for example when a radiologist has to judge and communicate the statistical meaning of a positive mammography screening. Many empirical studies have documented faulty inferences and even cognitive illusions among professionals of various disciplines (Hoffrage et al., 2000; Operskalski and Barbey, 2016). In the medical context, the consequences are particularly severe because many patients are mistakenly found diseased, which can entirely change their lives (Brewer et al., 2007; Gigerenzer et al., 2007; Salz et al., 2010; Wegwarth and Gigerenzer, 2013). Similarly, insufficient knowledge of statistics in general and incorrect Bayesian reasoning in particular can result in false convictions or acquittals made by juries in court, for example when they have to evaluate evidence based on a fragmentary DNA sample. These faults bear the risk of destroying innocent people's lives, too, as happened, for instance, in the famous case of Sally Clark (Schneps and Colmez, 2013; Barker, 2017).

Typically, the statistical information that the aforementioned professionals are confronted with is provided in probability format, that is, fractions or percentages describing the probability of a single event, for example the prevalence of breast cancer in the population. Generally, in situations where Bayesian inferences are necessary, three pieces of statistical information are given: the base rate (or a priori probability), sensitivity, and false alarm rate. Consider, for instance,

the heroin addiction problem (adapted from Gigerenzer and Hoffrage, 1995):

Heroin addiction problem (probability format):

> The probability of being addicted to heroin is 0.01% for a person randomly picked from a population *(base rate)*. If a randomly picked person from this population is addicted to heroin, the probability is 100% that he or she will have fresh needle pricks *(sensitivity)*. If a randomly picked person from this population is not addicted to heroin, the probability is 0.19% that he or she will still have fresh needle pricks *(false alarm rate)*. What is the probability that a randomly picked person from this population who has fresh needle pricks is addicted to heroin (*posterior probability*)?

With the help of Bayes' theorem, the corresponding posterior probability P(H|N), with H denoting "person is addicted to heroin" and N denoting "person has fresh needle pricks," can be calculated.

$$P(H|N) = \frac{P(N|H) \cdot P(H)}{P(N|H) \cdot P(H) + P(N|\neg H) \cdot P(\neg H)} = \frac{100\,\% \cdot 0.01\,\%}{100\,\% \cdot 0.01\,\% + 0.19\,\% \cdot 99.99\,\%} \approx 5\,\% \quad (1)$$

Given the probabilistic information (the low base rate, high sensitivity, and low false alarm rate), the result of only 5% seems astonishingly low to most people—professionals and laypeople alike. In fact, only very few—on average as few as 4% of the participants included in a comprehensive meta-analysis (McDowell and Jacobs, 2017)—are able to draw the correct inferences necessary to come to the right conclusion in such Bayesian tasks. The vast majority of people have difficulties, which can result in severe misjudgments.

The reasons for this poor performance in Bayesian reasoning are widely discussed. One explanation is the neglect of the base rate, which can be very low in many Bayesian situations (Tversky and Kahneman, 1974; Bar-Hillel, 1983). This leads to much greater estimates for the posterior probability, which is consistent with most people's intuition. Further reasons for the poor performance include participants neglecting the false alarm rate P(N|H) or confusing the false alarm rate with the posterior probability P(H|N) (Gigerenzer and Hoffrage, 1995) as well as participants overweighing the sensitivity (e.g., McCloy et al., 2007).

In order to prevent dangerous misjudgments due to faulty Bayesian inferences, the concept of *natural frequencies* has proven to be a powerful instrument (e.g., Gigerenzer and Hoffrage, 1995; Siegrist and Keller, 2011). Natural frequencies can be obtained by *natural sampling* (Kleiter, 1994) or, alternatively, by translating probabilities (e.g., "80%") into expressions consisting of two absolute frequencies (e.g., "80 out of 100"; for a discussion on the equivalence of natural frequencies and probabilities, see section Present Approach). Consider once again the heroin addiction example, this time, however, in natural frequency format:

> 10 out of 100,000 people from a given population are addicted to heroin. 10 out of 10 people who are addicted to heroin will have fresh needle pricks. 190 out of 99,990 people who are not addicted to heroin will nevertheless have fresh needle pricks. How many of the people from this population who have fresh needle pricks are addicted to heroin?

With the help of this format, significantly more people find the correct answer to the problem, which is 10 out of (10 + 190). As a consequence, performance rates in the frequency format typically increase to about 24% (McDowell and Jacobs, 2017). Errors due to base rate neglect as mentioned above occur less often with natural frequencies, since the base rate need not be attended to in the frequency version because it is already included in the information on the sensitivity and false alarm rate. Thus, Bayes' modified theorem containing natural frequencies yields the correct answer of "10 out of 200" in the heroin addiction problem based on a simpler computation:

$$P(H|N) = \frac{\#(N \cap H)}{\#(N)} = \frac{10}{10 + 190} = 5\% \quad (2)$$

More than 20 years of research have confirmed the benefit that comes with the concept of natural frequencies in Bayesian reasoning situations. Laypeople, students, professionals across various domains (e.g., medicine, law, and management), and even children perform significantly better when working on a Bayesian reasoning task that is presented in natural frequencies instead of probabilities (e.g., Wassner, 2004; Zhu and Gigerenzer, 2006; Hoffrage et al., 2015; Binder et al., 2018).

Additionally, various other factors are known to have an impact on performance in Bayesian reasoning tasks. Visualizations, for example tree diagrams (e.g., Yamagishi, 2003; Binder et al., 2018), unit squares (e.g., Böcherer-Linder and Eichler, 2017; Pfannkuch and Budgett, 2017), icon arrays (e.g., Brase, 2009, 2014) or roulette wheel diagrams (e.g., Yamagishi, 2003; Brase, 2014), have been shown to improve accuracies in Bayesian situations (for an exception, see, e.g., Micallef et al., 2012). An overview and categorization of visualizations that were used to boost performance in Bayesian situations is provided by Khan et al. (2015). Furthermore, individual differences of participants, particularly cognitive abilities such as numeracy, graphicacy, and spatial abilities, certainly have an impact on performance rates (e.g., Chapman and Liu, 2009; Brown et al., 2011; Micallef et al., 2012; Peters, 2012; Ottley et al., 2016). In addition, the specific numerical values for population size, base rate, sensitivity, and false alarm rate can influence accuracies (Schapira et al., 2001). Cognitive biases and judgment errors associated with different numerical information are, for example, size effect and distance effect (Moyer and Landauer, 1967). Finally, details of the representation and framing of the problem text can affect performance in Bayesian reasoning situations (Obrecht et al., 2012). Ottley et al. (2016), for example, were able to show that specific problem formulations (e.g., providing *all* numerical information in context of the task, that is, not only base rate, sensitivity, and false alarm rate but also the probability or frequency of their respective complement) influence accuracies significantly.

However, instead of contributing to the abundance of empirical studies replicating and discussing the beneficial effect

of natural frequencies or other factors (e.g., Hoffrage et al., 2002; Pighin et al., 2016; McDowell et al., 2018), in this article we will focus on the other side of the coin, that is, on the 76% of participants in these studies (on average in McDowell and Jacobs, 2017) who failed to solve Bayesian reasoning tasks with natural frequencies. Why can still on average only a quarter of participants solve the problem correctly, although the task is presented in the beneficial natural frequency format? Many psychological theories explain, discuss, and specify in detail if and why natural frequencies facilitate Bayesian inferences (e.g., the nested sets-hypothesis or the ecological rationality framework, see Gigerenzer and Hoffrage, 1999; Lewis and Keren, 1999; Mellers and McGraw, 1999; Girotto and Gonzalez, 2001, 2002; Hoffrage et al., 2002; Sloman et al., 2003; Barbey and Sloman, 2007; Pighin et al., 2016; McDowell et al., 2018) and how additional tools, such as visualizations, further increase their beneficial effect (e.g., Yamagishi, 2003; Brase, 2009, 2014; Spiegelhalter et al., 2011; Micallef et al., 2012; Garcia-Retamero and Hoffrage, 2013; Micallef, 2013; Ottley et al., 2016; Böcherer-Linder and Eichler, 2017). However, a satisfying answer to the question why only 24% of participants solve Bayesian reasoning problems in natural frequency format correctly has not yet been found.

PRESENT APPROACH

In order to explain why only 24% of participants draw correct Bayesian inferences when confronted with natural frequencies, in the present article we take one step back and switch our focus from *performance rates* to *cognitive processes*. In this respect, some important questions have not been addressed in detail so far: When given a Bayesian reasoning problem in frequency format, how do participants who fail to provide the correct answer approach the task? Where exactly do their calculations fail and why?

In order to gain a first impression of what participants might do when confronted with a task in natural frequency format, we checked the questionnaires from our previous studies on Bayesian reasoning and natural frequencies (e.g., Krauss et al., 1999; Binder et al., 2015). Interestingly, we revealed some instances where participants had not applied the given natural frequencies but had translated them back into probabilities. In order to explore this phenomenon in depth, we had a closer look on what students usually learn about Bayesian reasoning problems in their high school statistics classes.

Over the past two decades, statistics education has become an important column in German high school curricula. Here, just like in other countries, systematic calculation with probabilities has been in the center of teaching efforts. Alternative formats, such as natural frequencies, have despite the great amount of empirical research underpinning their benefits only played a minor role (cf. the American GAISE recommendations; Franklin et al., 2007). Even though there are some very recent efforts to implement the frequency concept in German curricula, for example in the new Bavarian high school curriculum for grade 10 (ISB, 2016), there still seems to be a tendency that this format is not accepted as equally mathematically valid as probabilities. This is supported by our impression from trainings for mathematics teachers that the concept of natural frequencies is not even familiar to most teachers. Furthermore, many schoolbooks tend to solve statistical tasks (not only Bayesian ones) with probability calculations, even when the task is presented in absolute frequencies (e.g., Freytag et al., 2008; Rach, 2018). Another observation we made based on a review of typical Bavarian school textbooks (Eisentraut et al., 2008; Freytag et al., 2008; Schmid et al., 2008) and workbooks (Sendner and Ruf-Oesterreicher, 2011; Reimann and Bichler, 2015) was that the more advanced students become in their high school career, the fewer statistical tasks are solved with natural frequencies by the respective textbooks. In conclusion, high school (and, consequently, university) students are a lot more familiar with probabilities than with natural frequencies due to their general (and sometimes even tertiary) statistical education. This implies that working with probabilities is a well-established strategy when it comes to solving statistical problems.

While in many situations people profit from such an established strategy, in some cases, however, a previously fixed mindset can block simpler ways to approaching a problem (Haager et al., 2014). This phenomenon lies at the center of prominent psychological theories on cognitive rigidity. Consider, for example, the so-called Einstellung or mental set effect (Luchins, 1942). When solving a problem, people often rigidly apply a previously learnt solution strategy while neglecting possibly important information that would allow an easier solution. Such an *Einstellung* or *mental set* can be developed through repeated training, enabling the person to quickly solve problems of the same structure (Schultz and Searleman, 2002; Ellis and Reingold, 2014; Haager et al., 2014). However, the downside of these mental sets is that they can make a person "blind" to simpler solutions or—in the worst case—unable to find a solution at all.

The most famous example for the Einstellung effect is Luchin's water jar experiment (1942; for more recent studies on the Einstellung effect in chess players and with anagram problems see, e.g., Bilalić et al., 2008; Ellis and Reingold, 2014). Participants in Luchin's study had to work out on paper how to obtain a certain volume of water using three empty jars of different sizes for measuring. The first five problems could all be solved by applying a relatively complicated strategy that was shown to the participants in an example problem. For the following five problems, a much simpler solution method was possible. However, the majority of participants kept using the complicated strategy they had previously learnt. Moreover, many of them could not solve the eighth problem at all, for which only the simple solution strategy was appropriate (Luchins, 1942).

Recent research has shown that even experts can be subject to the Einstellung effect (e.g., Bilalić et al., 2008). Thus, mental sets developed over a long period of time can also lead to the blocking of simple solutions (for a detailed discussion of different aspects of cognitive rigidity see Schultz and Searleman, 2002). The probability strategy, which German students deal with during their whole high school career, would be an example for such a mental set that is developed over time. So taken together, these psychological theories and the strong familiarity of students with probabilities hint toward a possible answer to the question what participants might wish to do when they are confronted

with a task in frequency format: They might try to represent the situation in the much more familiar probability format in order to be able to use established probabilities for their calculations.

Such an Einstellung toward calculating with probabilities instead of natural frequencies would take away all benefits that come with the frequency concept. Calculating with probabilities in a Bayesian context—even though the task is provided in frequency format—has the consequence that the intuitive natural frequency algorithm [formula (2)] is no longer available, the more complicated probability algorithm [formula (1)] has to be applied, and people are no longer able find the correct solution. Thus, the Einstellung effect might explain why on average three quarters of participants fail with natural frequencies. In the same line, we assume that it is very unlikely that people translate probabilities into natural frequencies when given a task in probability format—despite over 20 years of research on the beneficial effects of natural frequencies.

Here, the question might arise whether the two formats can actually be considered equivalent. In this respect, both mathematical and psychological aspects need to be addressed. First, we will shed light on the respective mathematical frameworks both formats operate in and to what extent these frameworks can be considered equivalent. Second, we will analyze the equivalence of probabilities and natural frequencies from a psychological viewpoint.

Even though the two formats seem to follow different rules, from a mathematical perspective they can be defined analogously. Weber (2016) showed that natural frequencies can be embedded in a theoretical framework that is isomorphic to a probability space, that is, the structure at the basis of probability theory can be constructed in a similar way for natural frequencies. Thus, all fundamental mathematical properties of probabilities, for example closure, commutativity, and associativity of their addition, can theoretically also be assigned to natural frequencies (for details, see Weber, 2016). Therefore, the two concepts can be considered equivalent, implying that natural frequencies are an information format just as mathematically valid as probabilities.

However, regardless of this theoretical equivalence of the two formats, a certain psychological uneasiness about the equivalence of natural frequencies and probabilities still seems to exist. It can be speculated that students who do not know about the mathematical framework of the frequency format might switch from natural frequencies to probabilities not only because they think that a probability algorithm is the only or the easiest way to solve the problem but also due to this subtle feeling of uneasiness, which stems from the assumption that natural frequencies are not a mathematically valid tool for solving Bayesian reasoning tasks. The latter implies that participants—even if they realize that a solution can be derived very easily by using natural frequencies—might think that a mathematically justified argumentation requires reasoning in terms of probabilities. All three assumptions (probabilities are the only, the easiest or the only allowed way) might trigger participants to rely on their Einstellung instead of actively using natural frequencies.

To be clear, we theoretically consider natural frequencies as a superordinate concept for both "expected" and "empirically sampled" frequencies. Expected frequencies constitute frequencies expected in the long run (cf. Hertwig et al., 2004; Spiegelhalter and Gage, 2015; case 2 in Woike et al., 2017) and are often used for problem formulations in natural frequency format. In contrast, empirically sampled frequencies are derived from a natural sampling process (cf. Kleiter, 1994; Fiedler et al., 2000; cases 1 and 3 in Woike et al., 2017; for a discussion of the two sub-concepts of natural frequencies, see also Hertwig et al., 2004; Spiegelhalter and Gage, 2015).

Of course, in the context of possibly switching between the two formats, besides the information format of the task, also the format in which the *question* is asked has to be taken into consideration (for a discussion on other details of textual problem representation, see, e.g., Ottley et al., 2016). It has to be noted that several studies (e.g., Cosmides and Tooby, 1996; Evans et al., 2000; Girotto and Gonzalez, 2001; Sirota et al., 2015) suggest that a question format that does not match the information format of the task reduces the natural frequency facilitation effect (Ayal and Beyth-Marom, 2014; Johnson and Tubau, 2015). However, only few studies directly test such incongruent problem and question formats (McDowell and Jacobs, 2017).

We also do not want to examine incongruent formats (or other factors mentioned above) systematically (e.g., in order to boost performance), but rather aim to implement a question format as neutral as possible that allows for both answer formats simultaneously. Our interest is to observe and analyze a substantial amount of participants for all four possible cases, namely those who stay with the given format (probability or natural frequency) and those who switch to the other format for their calculations, in order to learn from the respective cognitive processes about possible mechanisms underlying the choice of calculation format.

Since in our questionnaires from previous studies (Krauss et al., 1999; Binder et al., 2015), it was not always possible to judge which calculation format a participant applied, we will now explicitly ask participants to write down their solution algorithm in order to capture cognitive policies. Thus, in the present study we enter new research fields by investigating potential preferences in *calculation format*—when a problem introduction and question format as neutral as possible are given—that become visible by the way participants try to solve a given Bayesian task.

Our research questions are:

- Research question 1: Do participants show a general preference of the probability format over natural frequencies that becomes manifest in a strong tendency to

a) keep working with probabilities if a task is given in probability format, although a sample population is provided
b) even translate a task given in frequency format into probabilities, if the question allows for answers in both formats?

- Research question 2:

a) Regardless of the format in which the task is presented, do participants who work on this task actively using natural frequencies make more correct Bayesian inferences than participants who make their computations with probabilities?
b) If questions allow for answers in both formats, which factor predicts correct Bayesian inferences better—the format that the task is presented in (*presentation format*) or the format that participants actively use for their calculations (*calculation format*)?

Regarding research question 1, we hypothesized that participants do show a strong preference of probabilities over natural frequencies in both presentation formats. We further assumed that this preference has indeed a detrimental effect on performance in Bayesian reasoning tasks. With regard to research question 2, we therefore hypothesized that actively working with natural frequencies is a stronger predictor for correct inferences than the presentation format of a task.

EXPERIMENTAL STUDY

To examine these research questions, we conducted an empirical study with a first sample ($N = 114$) in 2016 (see section Participants). In the light of the current debate on the replication crisis (e.g., Open Science Collaboration, 2015), we decided to check the robustness of the results obtained with another sample ($N = 69$) with the same materials and design in 2017/2018. Three participants from the second sample were excluded from the analysis because they indicated that they had already participated in the first sample. Since we detected the same effects for both samples independently, we report the results for the combined sample of $N = 180$ (see section Results).

Method

Participants in our study had to work on two Bayesian reasoning tasks with different scenarios (heroin addiction problem and car accident problem, adapted from Gigerenzer and Hoffrage, 1995) and different numerical data (for design see **Table 1** and for problem wordings see **Table 2**). These two contexts were chosen since they are not as common as, for example, the famous mammography problem, and thus, the chance of a participant already knowing the task beforehand was small. Moreover, both problems refer to daily-life situations, so the participants were expected to have no difficulties understanding the scenarios. One of the two Bayesian problems was presented in probability format and the other one in natural frequency format. We systematically permuted the order of context as well as information format.

In typical natural frequency versions, the question reads "How many of the … have/are …?," often followed by a line "Answer: ____ out of ____." Note that we are interested in cognitive processes triggered purely by the *presentation format* and not by a provided question or answer format. Thus, in all natural frequency versions, we wanted to implement a question format that allows both for probability and for natural frequency answers. In order to be as neutral as possible, we decided to use questions for *proportions* (see **Tables 1**, **2**), which are a common question format in schoolbooks, too. The question "What is the proportion of people…" can be answered by, for example, "5%" or by "10 out of 200" and thus is settled in between probabilities and natural frequencies.

TABLE 1 | Design of the implemented problem versions.

		Context	
		Heroin addiction problem	**Car accident problem**
Presentation format	Probabilities	• Introduction: sample provided • Presentation format of the task: probabilities • Question format: probabilities • Visualization presented or to be constructed	• Introduction: sample provided • Presentation format of the task: probabilities • Question format: probabilities • Visualization presented or to be constructed
	Natural frequencies	• Introduction: sample provided • Presentation format of the task: natural frequencies • Question format: proportions • Visualization presented or to be constructed	• Introduction: sample provided • Presentation format of the task: natural frequencies • Question format: proportions • Visualization presented or to be constructed

In the probability versions, formulating a neutral question is rather difficult because a proportion usually refers to a concrete sample. Thus, instead of making the question format as neutral as possible, we decided to provide the participants already in the introduction with a sample population that the probabilities could be referred to (e.g., "On the internet, you find the following information for a sample of 100,000 people"). Thereby, we again allowed for both calculation formats. While in natural frequency versions the option for probability answers lies in the neutral question format, a possible natural frequency answer in probability versions was opened up by providing a concrete sample in the beginning of the task. It is important to note that we did not primarily want to compare performances by *presentation format* (which would just be a replication of many other studies) but by *calculation format*, so a total parallelization of the task versions was neither necessary nor the optimal design for our research questions.

Because Bayesian reasoning tasks in German schoolbooks are usually presented with tree diagrams (Binder et al., 2015), after the question, we either asked for the construction of a tree diagram (in the first task) or presented a tree diagram (in the second task). The aim here was to present stimuli that are as ecologically valid as possible [with respect to (German) teaching contexts both in school and in university] and that provide the option to switch between the two formats. Both at school and at university level, 2 × 2-tables and tree diagrams are most commonly used for teaching Bayesian reasoning, whereas alternative visualizations (unit squares, icon arrays, etc.) are usually omitted. Since both 2 × 2-tables and tree diagrams allow

TABLE 2 | Problem formulations.

	Heroin addiction problem		Car accident problem	
	Probability version	**Natural frequency version**	**Probability version**	**Natural frequency version**
Introduction	Imagine that you randomly meet a person with fresh needle pricks in the street. You are interested in whether this person is addicted to heroin. On the internet, you find the following information for a sample of 100,000 people:		Imagine you see a drunken person getting behind the wheel of his or her car after a party. You are interested in the risk of a car accident caused by this person. On the internet, you find the following information for a sample of 10,000 drivers:	
Statistical information	The probability that one of these people is addicted to heroin is 0.01%.	10 out of 100,000 people are addicted to heroin.	The probability that one of these drivers will cause an accident is 1%.	100 out of 10,000 drivers cause an accident.
	If one of these people is addicted to heroin, the probability is 100% that he or she will have fresh needle pricks.	10 out of 10 people who are addicted to heroin will have fresh needle pricks.	If one of these drivers causes an accident, the probability is 55% that he or she is drunk.	55 out of 100 drivers who cause an accident are drunk.
	If one of these people is not addicted to heroin, the probability is 0.19% that he or she will nevertheless have fresh needle pricks.	190 out of 99,990 people who are not addicted to heroin will nevertheless have fresh needle pricks.	If one of these drivers does not cause an accident, the probability is 5% that he or she is nevertheless drunk.	500 out of 9,900 drivers who do not cause an accident are nevertheless drunk.
Question	What is the probability that one of these people is addicted to heroin, if he or she has fresh needle pricks?	Of the people who have fresh needle pricks, what is the proportion of them addicted to heroin?	What is the probability that one of these drivers causes an accident, if he or she is drunk?	Of the drivers who are drunk, what is the proportion of them causing an accident?
Visual aid	• First task: construct a tree diagram • Second task: consider a presented tree diagram	• First task: construct a tree diagram • Second task: consider a presented tree diagram	• First task: construct a tree diagram • Second task: consider a presented tree diagram	• First task: construct a tree diagram • Second task: consider a presented tree diagram
Prompt	"Please write down your calculations!"			

for switching between the two formats (unlike, e.g., icon arrays) and since tree diagrams but not 2 × 2-tables can be directly equipped with *conditional* probabilities, only tree diagrams remained as visualizations suitable for our study. By using the latter, our hope was to exploratively shed light on whether a tree diagram might influence participants' choice of calculation format, for example by making the given presentation format more salient (for tree diagrams equipped with probabilities or natural frequencies in the heroin addiction problem see **Figure 1**). In sum, rather than systematically varying specific factors (or boosting performance), we wanted (1) to know how participants reason with the materials usually presented in German schools and universities, and (2) to observe a substantial number of people switching or staying with the presentation format in order to analyze their respective reasoning processes. For the same reasons, we implemented standard problem wordings.

Since participants were explicitly asked to write down all calculations they made in order to solve the task, we were able to judge precisely and systematically which format they used for their calculations (see **Supplementary Table 2**; also see section Coding).

The paper and pencil questionnaire contained a short information paper on the study and some general questions, for example on participants' age or study program, as well as the two tasks. Before participants were allowed to start with the second task, they had to hand in their solution for the first task. Participants were allowed to use a pocket calculator that was provided along with the questionnaire. There was no time limit; on average, participants took approximately 5 min to complete the demographic items and 25 min for both tasks.

Coding

The normatively correct solutions of the problems were 5% (or 10 out of 200) for the heroin addiction problem and 9.9% (or 55 out of 555) for the car accident problem (the results differ marginally if the task was presented in natural frequencies as opposed to probabilities, e.g., exactly 10% in the car accident probability version vs. 9.9% in the car accident frequency version). In order to guarantee maximum objectivity for classifying the answers as "correct Bayesian inference" or "incorrect Bayesian inference" and also for deciding whether either a probability algorithm or a frequency algorithm had been applied, we used strict coding guidelines (see **Supplementary Table 1**), which were applied by all coders. Since we were especially interested in whether participants used the correct *algorithm* for solving the task, mere calculation or rounding errors were neglected, resulting in answers that were classified as "correct Bayesian inference" even though the mathematical result was not entirely correct. In the same line, answers that appeared mathematically correct at first glance were classified as "incorrect Bayesian inference" if the result was just incidentally correct, but a wrong algorithm was applied (this rarely happened).

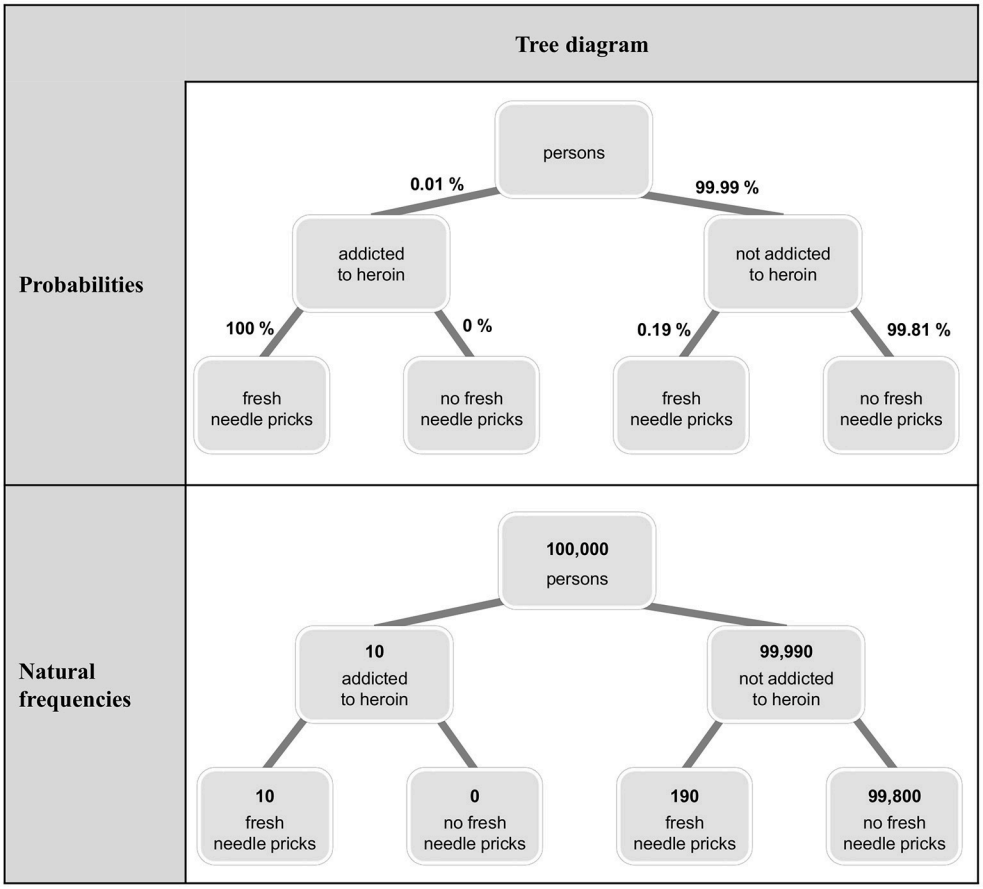


FIGURE 1 | Tree diagrams visualizing the heroin addiction problem equipped with probabilities and natural frequencies.

Furthermore, we focused on the cognitive processes underlying each response when determining the "calculation format" of an answer. This cognitive process was measured by analyzing the exact calculations each participant wrote down to come to a solution. When a participant used probabilities (or natural frequencies) only, we classified the solution as "calculated with probabilities" (or natural frequencies, respectively). When both formats were clearly visible in the calculations, we classified the answer according to whether the participant used probabilities or natural frequencies for the *crucial step* in the calculation process, that is, the computation of the denominator in Bayes' formula, as can be seen in equations (1) and (2). Thus, the decisive factor in such unclear cases was the *addition* of two absolute numbers (in favor of a frequency algorithm) or the *multiplication* of probabilities (in favor of a probability algorithm, respectively). If, for example, in the heroin addiction problem a participant used both formats for his or her calculations, but *added* two absolute numbers (e.g., $10 + 190$) to obtain the denominator in (2), the answer was classified as "calculated with natural frequencies". If, on the other hand, a participant used both formats, but *multiplied* two probabilities (e.g., $0.01 \times 100\%$) like in (1) to obtain the respective probabilities for the numerator or the denominator, we classified the answer as "calculated with probabilities" (no participant added frequencies *and* multiplied probabilities).

Two raters coded 21% of all inferences independently according to the coding guidelines (see **Supplementary Tables 1, 2**). Since in 100% of all cases the correctness was rated in congruence (Cohen's $\kappa = 1$; Cohen, 1960), and the calculation format was classified identically in 97% of all cases (Cohen's $\kappa = 0.95$), the remaining inferences were rated by one coder.

Participants

We recruited $N = 114$ students from the University of Regensburg (Bavaria) in summer 2016, and $N = 69$ in winter 2017/2018 (three of which were excluded from the analysis since they had already participated in the study in 2016). Most of these students were enrolled in a teaching math program ($N = 147$), while some of them studied economic information technology, so a certain level of mathematics competency among the participants can be assumed (see also section Discussion). They were at different stages of their studies (most of them in their first two years) and their age ranged from 18 to 38, with an average of 22 years. Out of the total of $N = 180$ participants, 121 were female. Since each participant worked on two tasks, we

obtained a total of 360 Bayesian inferences including participants' detailed solution algorithms.

The study was carried out in accordance with the University Research Ethics Standards. Participants were informed that the study was voluntary and anonymous, and no incentives were paid. Participants were asked to give their written informed consent to participate in the study in advance. Thereupon, two students refrained from participating.

RESULTS

In the following, we report the results for the combined sample of $N = 180$ participants, but all detected effects also hold for both the original ($N = 114$) and the replication sample ($N = 66$) independently. As far as our first research question is concerned, the results indeed show a strong preference of participants for calculating with probabilities in both contexts. This is illustrated by **Figure 2**, where, for example, P→ F denotes participants who were provided with a task in probability format but calculated with natural frequencies. On the one hand, when presented with a task in natural frequency format (second and fourth bars of **Figure 2**), almost half of participants (49%) nevertheless chose to apply probabilities for their calculations, although the neutral question explicitly allowed for answers in both formats. On the other hand, when they faced a probability version of a task (first and third bars of **Figure 2**), only 18% across both contexts chose to translate the problem into natural frequencies—despite the explicitly given sample population in the introduction. Taken together, according to our design natural frequencies represented the preferred calculation format in only about one third (34%) of all 360 Bayesian tasks although 50% of all tasks were presented in natural frequency format.

While **Figure 2** does not yet display performances, **Figure 3** shows performance rates in the resulting four combinations of presentation format and calculation format (P→P, P→F, F→F, F→P) for both problem contexts. It becomes clear that when natural frequencies were actively used for the calculations, performance rates were significantly higher than when probabilities were applied. Remarkably, in our design this holds true almost regardless of the presentation format: For both problems, the patterns look very similar for the two presentation formats. The performance in both problems obviously mainly depends on the calculation format, but only to a small amount on the presentation format. In the heroin addiction problem, the difference between both calculation formats is especially pronounced. The highest performance was detected when both variables *presentation format* and *calculation format* were natural frequencies (61% correct responses), descriptively followed by probability tasks that were worked on with frequencies (53% correct responses). In the two other cases (when participants calculated with probabilities), performance rates were considerably lower (13% if the presentation format was probabilities and 9% if the presentation format was natural frequencies).

In general, the beneficial effect of presenting natural frequencies was replicated by our study. While 20% of the Bayesian tasks in probability format were solved correctly across both contexts, the performance rate for the tasks presented in frequency format was 36% (see **Table 3**). Compared to McDowell and Jacobs (2017), both of these numbers seem rather high. An explanation might lie within our sample: more than 80% of participants were enrolled in a mathematics education program and might therefore have comparably high numeracy, enabling them to perform above average in math tasks (for an analysis of participants' individual differences and switching behavior depending on their cognitive abilities, see below). Note that we also found context effects (36% correct responses in the heroin context vs. 20% correct inferences in the car accident context).

In order to separate the effects of presentation format and calculation format, we ran a generalized linear mixed model (GLMM) with a logistic link function. Here, we specified probabilities (both as presentation format and as calculation format) as reference category and included the possible explanatory factors "presentation format," "calculation format" (via dummy coding), and the interaction term of presentation format and calculation format to predict the probability of a correct Bayesian inference in our design.

According to the results of the generalized linear mixed model, the unstandardized regression coefficient for solving a task that was both presented and calculated in probability format was significant ($b_0 = -7.03$, SE $= 1.32$, $z = -5.32$, $p < 0.001$), showing large inter-individual differences (for a discussion of these results, see below). The (unstandardized) regression coefficient for the *presentation format* was non-significant ($b_1 = -3.04$, SE $= 2.00$, $z = -1.52$, $p = 0.13$), whereas the *calculation format* showed a significant regression coefficient ($b_2 = 9.85$, SE $= 3.85$, $z = 2.56$, $p = 0.01$). Finally, the interaction of presentation format and calculation format yielded another significant regression coefficient ($b_3 = 4.85$, SE $= 2.22$, $z = 2.19$, $p = 0.03$), indicating that calculating with natural frequencies increases performance even more when the task is also formulated in natural frequency format (i.e., when the absolute numbers for the frequency algorithm can be directly taken from the problem wording).

The strong differences of individual competencies lead to extreme (unstandardized) regression coefficients in the model. However, a generalized linear model (neglecting inter-individual differences) estimated regression coefficients that—converted into probabilities via the logistic link function—exactly replicated the performance rates found in our data. This is because the GLMM accounts for these large differences in performances by estimating large inter-individual differences between the participants, as the intercepts (denoting the performances when presentation and calculation format was probabilities) were allowed to vary freely between participants. The substantial influence of the inter-individual differences also becomes apparent when inspecting the model fit: Whereas 6.5% of the variance is explained by the fixed GLMM regression coefficients (marginal $R^2 = 0.065$), the inter-individual differences and the fixed regression coefficients together explain 68.5% of the variance (conditional $R^2 = 0.685$). However notably, despite the large inter-individual differences, the influence of the fixed effects on the results was clear and strong.

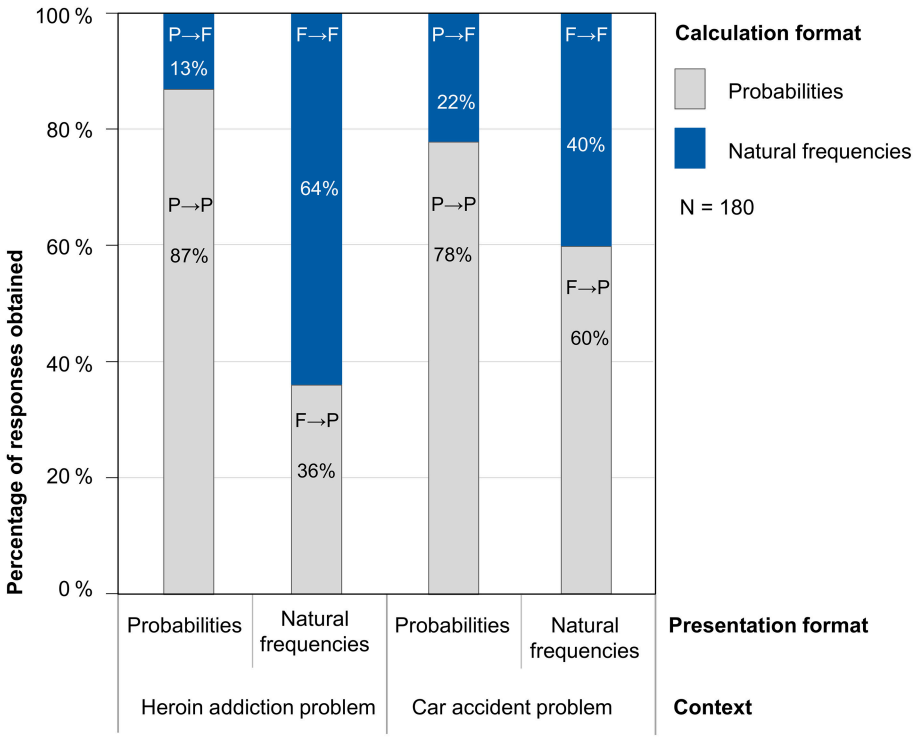


FIGURE 2 | Calculation format by presentation format and context.

TABLE 3 | Percentage of correct Bayesian inferences by context and presentation format (independent of calculation format).

Presentation format	Context		Average
	Heroin addiction problem	Car accident problem	
Probabilities	22% ($n = 92$ inferences)	19% ($n = 89$ inferences)	20% ($n = 181$ inferences)
Natural frequencies	51% ($n = 87$ inferences)	22% ($n = 92$ inferences)	36% ($n = 179$ inferences)

Although we did not explicitly collect data about participants' cognitive abilities (e.g., numeracy, spatial and graphical literacy), these inter-individual differences suggested a closer analysis of our data with this respect. Indeed, we found significant differences in performance especially between two subgroups of our sample: The $N = 42$ mathematics education students aspiring to teach at the academic school track of the German school system (*Gymnasial students*) outperformed the other $N = 138$ participants significantly (50% correct inferences vs. 21%; $t(358) = 5.294$, $p < 0.001$). We assume that this difference is due to the higher numerical, spatial, and graphical abilities of the first group, since they generally outperform the other mathematics education students in mathematics exams or mathematical knowledge tests (e.g., Krauss et al., 2008; see also Lindl and Krauss, 2017, Table 5, p. 396). Moreover, the Gymnasial students receive a considerably more thorough education in mathematics through their study program than the rest of our participants. However interestingly, these differences in cognitive abilities did not have any influence on calculation format preferences. Both subgroups tended in a similar way to prefer using probabilities over natural frequencies for their calculations (32% of Gymnasial students' solutions were based on a frequency algorithm, whereas 35% of the other participants calculated with natural frequencies; $t(358) = -0.506$; $p = 0.613$). As a consequence, although an overall shift of performances might be expected depending on participants' cognitive abilities and education, we assume a certain generalizability of our results across varying abilities and education levels regarding the switching rates (cf. section Discussion).

By examining exploratively participants' reactions on a presented tree diagram, we revealed several instances where the participants had added probabilities to the branches of a tree diagram originally presented with natural frequencies in the nodes. Conversely, only few of the participants equipped a tree diagram that was originally presented in probability format with natural frequencies. When the participants had to construct actively a tree diagram visualizing the textual problem, we detected some instances where already before the diagram was drawn, participants had switched in their calculation format (in both directions: from natural frequencies to probabilities and vice versa). Therefore, some participants translated the presentation format into their calculation format right at the beginning of their problem solution process. However, since we did not

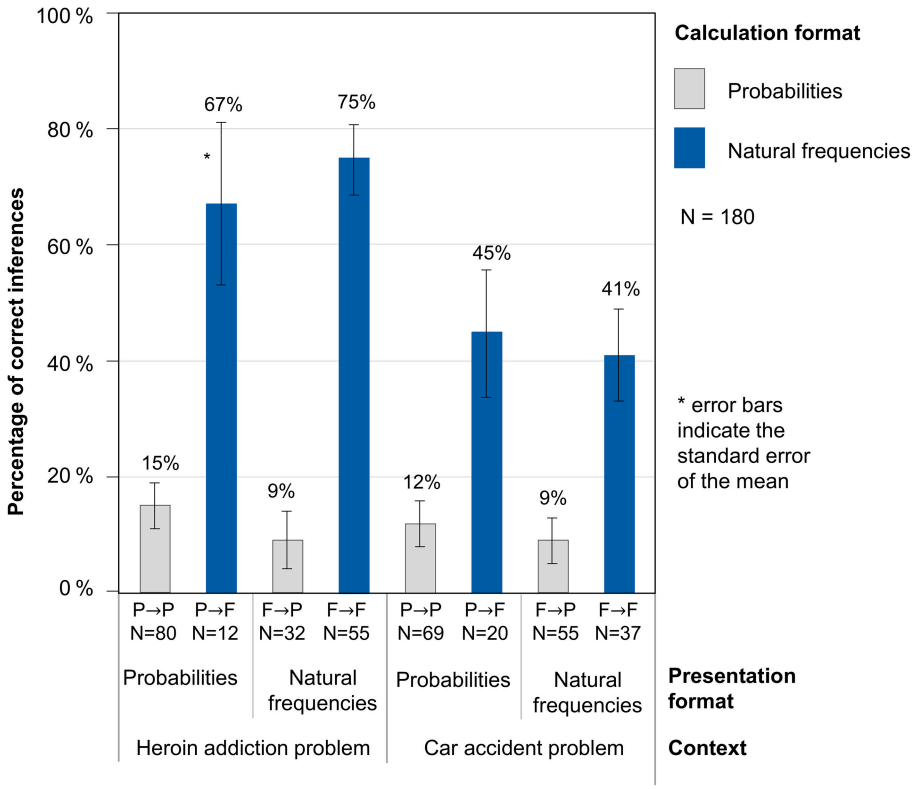


FIGURE 3 | Percentages of correct inferences dependent on the presentation and calculation format in both problems.

systematically test versions without a visualization clue, these findings have to be considered only explorative hints concerning possible cognitive mechanisms that might lead participants to stay with a certain format or to switch from one to the other. These mechanisms will have to be addressed more closely in future research.

DISCUSSION

In an empirical study with $N = 180$ students from the University of Regensburg, we found that the majority of participants do not actively use natural frequencies in Bayesian reasoning tasks. Even if the task is presented in the intuitive natural frequency format (with a neutral question asking for proportions), about half of the participants still prefer calculating with probabilities instead. Therefore, and since the "standardized" probability format is the "sine qua non" in probability theory, the results of our study reveal the Einstellung effect in Bayesian reasoning situations (Luchins, 1942; Luchins and Luchins, 1959; McCloy et al., 2007). We speculate that such an Einstellung might be enhanced by the still widespread idea that natural frequencies are not "mathematically correct" enough to actually work with in high school and university contexts. As a consequence, participants who might actually notice a possible solution of the Bayesian reasoning task based on a frequency algorithm might still rely on probabilities due to a certain kind of "phobia" to use natural frequencies for their calculations (for a discussion on the impact of affect on overcoming fixed mindsets, see Haager et al., 2014)—despite the ever-growing body of research pointing to the beneficial effects of the frequency concept (e.g., Gigerenzer and Hoffrage, 1995; Barbey and Sloman, 2007; Micallef et al., 2012; Obrecht et al., 2012; Ottley et al., 2016; McDowell and Jacobs, 2017).

Although with our study, we cannot ultimately decide whether the Einstellung effect or this kind of "phobia" lies at the heart of participants' switching back to probabilities, we want to emphasize that both formats are mathematically equivalent in the sense that they can be defined analogously with the same properties and structure. Whatever the case may be, since recent efforts to implement natural frequencies in high school and university curricula appear not to be enough to make people actively take advantage of their benefits, we vouch for an even stronger implementation of the natural frequency concept in secondary education (especially in the higher grades), tertiary education, and in teacher training.

The Einstellung toward preferring probabilities has a negative impact on performance rates: participants working with probabilities perform significantly worse than those who apply natural frequencies for their calculations. Moreover, at least in our design, the calculation format is an even stronger predictor for performance than the presentation format that previous

research has mainly concentrated on (e.g., Barbey and Sloman, 2007; Siegrist and Keller, 2011). This suggests that participants who translate natural frequencies into probabilities follow a path that is disadvantageous in two respects: First, they choose the unintuitive probability over the natural frequency format, and second, they are prone to make further mistakes due to translation errors (that we did not explicitly consider in our study). Interestingly, a few participants (18%) did translate probabilities into natural frequencies. This suggests that at least a small minority is to some extent familiar with the natural frequency concept. These participants profit indeed from calculating with natural frequencies since their performance rates increased substantially compared to performances of participants who stay with probabilities (13 vs. 53% across both implemented contexts). This tendency is a first sign that natural frequencies might become an established solution strategy for Bayesian reasoning tasks.

It has to be noted that our sample consisted of university students entirely. Since their mindsets and cognitive abilities (especially numeracy as well as graphical and spatial literacy) probably differ from the general population (Micallef et al., 2012), a different sample might, of course, yield different performance rates. However, we assume that even though the total population might generally perform worse than our sample, those using natural frequencies for their calculations will still outperform those who resort to probabilities. In the same way, we would expect an overall shift of performance rates depending on item difficulty or wording (for factors determining the difficulty of Bayesian reasoning tasks as well as for different problem wordings, see, e.g., Ottley et al., 2016), but we assume relative consistency with respect to format preferences across different Bayesian reasoning tasks. Future research might investigate in detail whether our results indicating an Einstellung effect in Bayesian reasoning situations hold also true when individual differences and item difficulty are systematically controlled.

The context effects in our study in favor of the heroin addiction problem could be explained by having a closer look at the question formulation in the car accident problem. Here, the two relative clauses in the frequency version (see **Table 2**) demand higher verbal processing abilities and thus make the question harder to understand compared to the frequency question in the heroin addiction problem (only one relative clause, see **Table 2**). Consequently, the heroin addiction problem presented with natural frequencies yields significantly higher performance rates than the respective version of the car accident problem (51% correct inferences vs. 22%; see **Table 3**). Moreover, coding in our study was fairly complex (see **Supplementary Tables 1, 2**), even though we obtained interrater reliability scores of $\kappa = 1$ for the correctness of a Bayesian inference and of $\kappa = 0.95$ (Cohen, 1960) for determining the calculation format. In addition, we focused only on the correct algorithm applied for classifying an answer as "correct" (see **Supplementary Table 1**). Thus, we did not concentrate on calculation errors, including those that resulted from translating an information format into the other one. Therefore, we did not systematically detect translation errors dependent on the respective presentation format, in particular. This, however, is a conservative approach, since we assume that more people make mistakes when translating frequencies into probabilities than vice versa.

Furthermore, in an explorative analysis, we detected several instances where the participants had equipped a presented frequency tree diagram with probabilities, suggesting that such a visualization does not prevent the participants from switching from the natural frequency to the probability format for their calculations. We speculate that even the opposite is the case: Since students are familiar with probability tree diagrams but not so much with frequency tree diagrams from their high school careers, the sight of a tree diagram (even though it is equipped with natural frequencies) might trigger their memories of the familiar probability trees and might thus provoke them to fill the diagram with probabilities. Moreover, many participants equipped the tree diagram they had been asked to draw with their chosen calculation format—even if the latter differed from the presentation format. This suggests that the participants tend to decide on their calculation format right at the beginning of their solution process. We thus speculate that the exact moment of the format switch lies immediately after (or even at the same time as) reading the task. Therefore, further research might investigate systematically when exactly people decide on the format they want to use for their calculations and if people possibly alter their decision during the solution process. In addition, it would be interesting to determine whether presenting a visualization such as a tree diagram or actively constructing one enhances or diminishes the Einstellung effect in Bayesian reasoning tasks (e.g., by systematically comparing versions with and without visualization)—and, more generally, whether visualizations affect the calculation format at all.

The question remains open to what extent natural frequencies should be implemented in statistics education, since they can only be used in specific situations (e.g., in Bayesian reasoning problems or tasks where cumulative risk judgment is necessary; see McCloy et al., 2007). We suggest that natural frequencies be taught already at a young age to establish the concept over a longer period of time. When—at a later stage—the focus is shifted more and more to probabilities, a permanent interplay between the two formats seems reasonable. By using natural frequencies to illustrate, for example, the multiplication rule or Bayes' theorem, students can understand the two coexisting formats as equally legitimate representations for the underlying concept of uncertainty. Here, natural frequencies can be used to eliminate typical errors, to make difficult problems more understandable, and to prevent cognitive illusions. When probabilities are presented simultaneously, the connection between the two formats might become more apparent and a deeper understanding of the concept of uncertainty might be achieved. In this respect, future work, for example systematic training studies (cf. Sedlmeier and Gigerenzer, 2001), needs to determine the most successful ways to incorporate natural frequencies in statistics education at secondary and tertiary level in order to overcome the Einstellung effect.

Future research on this topic might also investigate in more detail how much current teachers already know about the frequency concept in order to decide if natural frequencies indeed need a stronger focus in teacher training as we suggest. This could, for example, be realized by systematic teacher

interviews. Moreover, future research might address empirically the cognitive mechanisms underlying the Einstellung effect as detected by our study, that is, whether participants assume that a probability algorithm is (a) the only way, (b) the easiest way, or (c) due to a feeling of uneasiness with the frequency concept the only mathematically allowed way to approach the Bayesian problem. Here, qualitative methods such as student interviews might be a valuable tool to clarify situation-specific causes of the Einstellung effect. Finally, it would be interesting to determine effective methods (e.g., visualizations or hints in the problem wording) to prevent people from falling back into probabilities in Bayesian reasoning tasks.

DATA AVAILABILITY STATEMENT

The dataset generated can be found on https://epub.uni-regensburg.de/37693/.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of University Research Ethics Standards, University of Regensburg. The protocol was approved by the University of Regensburg. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We want to thank all participants of our study for contributing to our research project. We further thank Sven Hilbert for his statistical advice. This work was supported by the German Research Foundation (DFG) within the funding program Open Access Publishing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01833/full#supplementary-material

REFERENCES

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, 241–297. doi: 10.1017/S0140525X07001653

Bar-Hillel, M. (1983). The base rate fallacy controversy. *Adv. Psychol.*16, 39–61. doi: 10.1016/S0166-4115(08)62193-7

Barker, M. J. (2017). Connecting applied and theoretical Bayesian epistemology: data relevance, pragmatics, and the legal case of Sally Clark. *J. Appl. Philos.* 34, 242–262. doi: 10.1111/japp.12181

Bilalić, M., McLeod, P., and Gobet, F. (2008). Why good thoughts block better ones: the mechanism of the pernicious Einstellung (set) effect. *Cognition* 108, 652–661. doi: 10.1016/j.cognition.2008.05.005

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

Binder, K., Krauss, S., Bruckmaier, G., and Marienhagen (2018). Visualizing the Bayesian 2-test case: the effect of tree diagrams on medical decision making. *PLoS ONE* 13:e0195029 . doi: 10.1371/journal.pone.0195029

Böcherer-Linder, K., and Eichler, A. (2017). The impact of visualizing nested sets. An empirical study on tree diagrams and unit squares. *Front. Psychol.* 7:241. doi: 10.3389/fpsyg.2016.02026

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

Brase, G. (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

Brewer, N. T., Salz, T., and Lillie, S. E. (2007). Systematic review: the long-term effects of false-positive mammograms. *Ann. Intern. Med.* 146, 502–510. doi: 10.7326/0003-4819-146-7-200704030-00006

Brown, S. M., Culver, J. O., Osann, K. E., MacDonald, D. J., Sand, S., Thornton, A. A., et al. (2011). Health litereacy, numeracy, and interpretation of graphical breast cancer risk estimates. *Patient Educ. Couns.* 83, 92–98. doi: 10.1016/j.pec.2010.04.027

Chapman, G. B., and Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgm. Decis. Mak.* 4:34.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

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

Eisentraut, F., Ernst, S., Keck, K., Leeb, P., Schätz, U., Steuer, H., et al. and Schätz, R. (2008). *Delta 10 – Mathematik für Gymnasien [Delta 10 – Mathematics for the Academic School Track]*. Bamberg: CC Buchners Verlag.

Ellis, J. J., and Reingold, E. M. (2014). The Einstellung effect in anagram problem solving: evidence from eye movements. *Front. Psychol.* 5:679. doi: 10.3389/fpsyg.2014.00679

Evans, J. S. B. T., 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

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. General* 129, 399–418. doi: 10.1037//0096-3445.129.3.399

Franklin, C., Horton, N., Kader, G., Moreno, J., Murphy, M., Snider, V., et al. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report – A pre-K-12 Curriculum Framework*. Alexandria, VA: American Statistical Association. Available Online at: www.amstat.org/education/gaise

Freytag, C., Herz, A., Kammermeyer, F., Kurz, K., Peteranderl, M., Schmähling, R., et al. (2008). *Fokus Mathematik 10 Gymnasium Bayern [Focus on Mathematics 10 for the Bavarian Academic School Track]*. Berlin: Cornelsen Verlag.

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., 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

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/S00100277(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

Haager, J. S., Kuhbandner, C., and Pekrun, R. (2014). Overcoming fixed mindsets: the role of affect. *Cogn. Emot.* 28, 756–767. doi: 10.1080/02699931.2013.851645

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

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., 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

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

ISB, (2016). *Staatsinstitut für Schulqualität und Bildungsforschung LehrplanPLUS Gymnasium Mathematik 10 [Curriculum for year 10 of the Bavarian academic school track].* Available online at http://www.lehrplanplus.bayern.de/fachlehrplan/gymnasium/10/mathematik. (Accessed 18 July, 2018). [ISB] (Ed.).

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

Khan, A., Breslav, S., Glueck, M., and Hornbæk, K. (2015). Benefits of visualization in the mammography problem. *Int. J. Hum. Comput. Stud.* 83, 94–113. doi: 10.1016/j.ijhcs.2015.07.001

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.

Krauss, S., Baumert, J., and Blum, W. (2008). Secondary mathematics Teachers' pedagogical content knowledge and content knowledge: validation of the COACTIV constructs. *Int. J. Math. Educ.* 40, 873–892. doi: 10.1007/s11858-008-0141-9

Krauss, S., Martignon, L., and Hoffrage, U. (1999). "Simplifying Bayesian Inference: the General Case," in *Model-based Reasoning in Scientific Discovery*, ed N. E. A. Magnani (New York, NY: Kluwer Academic/Plenum Publishers), 165–179.

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

Lindl, A., and Krauss, S. (2017). "Transdisziplinäre Perspektiven auf domänenspezifische Lehrerkompetenzen. Eine Metaanalyse zentraler Resultate der Forschungsprojektes FALKO [Transdisciplinary perspectives on domain specific teacher competences. A meta-analysis of central results of the FALKO research project]," in *FALKO: Fachspezifische Lehrerkompetenzen. Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik [FALKO: Subject specific teacher competences. Conception of professional knowledge test in the subjects German, English, Latin, Physics, Musical Education, Evangelical Religious Education, and Pedagogy]*, eds S. Krauss, A. Lindl, A. Schilcher, M. Fricke, A. Göhring, B. Hofmann, P. Kirchhoff, and R. H. Mulder, (Münster: Waxmann), 381–438.

Luchins, A. S. (1942). Mechanization in problem solving: the effect of einstellung. *Psychol. Monogr.* 54, 1–95. doi: 10.1037/h0093502

Luchins, A. S., and Luchins, E. H. (1959). *Rigidity of Behavior: A Variational Approach to the Effect of Einstellung*. Eugene, OR: University of Oregon Books.

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

McDowell, M., Galesic, M., and Gigerenzer, G. (2018). Natural frequencies do foster public understanding of medical tests: comment on Pighin, Gonzalez, Savadori and Girotto (2016). *Medical Decis. Making.* 38, 390–399. doi: 10.1177/0272989X18754508

McDowell, M., and Jacobs, P. (2017). Meta-Analysis of the Effect of Natural Frequencies on Bayesian Reasoning. *Psychol. Bull.* 143, 1273–1312. doi: 10.1037/bul0000126

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. (2013). *Visualizing Set Relations and Cardinalities Using Venn and Euler Diagrams*. University of Kent. Dissertation.

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

Moyer, R. S., and Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature* 215, 1519–1520. doi: 10.1038/2151519a0

Obrecht, N. A., Anderson, B., Schulkin, J., and Chapman, G. B. (2012). Retrospective frequency formats promote consistent experience-based Bayesian judgments. *Appl. Cogn. Psychol.* 26, 436–440. doi: 10.1002/acp.2816

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349, 1–8. doi: 10.1126/science.aac4716

Operskalski, J. T., and Barbey, A. K. (2016). Risk literacy in medical decision-making. *Science* 352, 413–414. doi: 10.1126/science.aaf7966

Ottley, A., Peck, E. M., Harrison, L. T., Afergan, D., Ziemkiewicz, C., Taylor, H. A., et al. (2016). Improving Bayesian reasoning: the effects of phrasing, visualization, and spatial ability. *IEEE Trans. Vis. Comput. Graph.* 22, 529–538. doi: 10.1109/TVCG.2015.2467758

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

Pfannkuch, M., and Budgett, S. (2017). Reasoning from an eikosogram: an exploratory study. *Int. J. Res. Undergraduate Math. Educ.* 3, 283–310. doi: 10.1007/s40753-016-0043-0

Pighin, S., Gonzalez, M., Savadori, L., and Girotto, V. (2016). Natural frequencies do not foster public understanding of medical test results. *Medical Decision Making* 36, 686–691. doi: 10.1177/0272989X16640785

Rach, S. (2018). Visualisierungen bedingter Wahrscheinlichkeiten – Präferenzen von Schülerinnen und Schülern [Visualizations of conditional probabilities – preferences of students]. *Mathemat. Didact.* 41, 1–18.

Reimann, S., and Bichler, E. (2015). *Abitur 2016: Original-Prüfungsaufgaben mit Lösungen – Gymnasium Bayern Mathematik [Final secondary-school examinations 2016: Original mathematics exam tasks with solutions – Bavarian academic school track]*. Hallbergmoos: Stark Verlag.

Salz, T., Richman, A. R., and Brewer, N. T. (2010). Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. *Psycho-Oncol.* 19, 1026–1034. doi: 10.1002/pon.1676

Schapira, M. M., Nattinger, A. B., and McHorney, C. A. (2001). Frequency or probability? A qualitative study of risk communication formats used in health care. *Med. Decis. Making* 21, 459–467. doi: 10.1177/0272989X0102100604

Schmid, A., Weidig, I., Götz, H., Herbst, M., Kestler, C., Kosuch, H., et al. (2008). *Lambacher Schweizer 10 – Mathematik für Gymnasien Bayern [Lambacher Schweizer 10 – Mathematics for the Bavarian academic school track]*. Stuttgart: Ernst Klett Verlag.

Schneps, L., and Colmez, C. (2013). *Math on trial: How Numbers Get Used and Abused in the Courtroom*. New York, NY: Basic Books.

Schultz, P. W., and Searleman, A. (2002). Rigidity of thought and behavior: 100 years of research. *Genet. Soc. Gen. Psychol. Monogr.* 128, 165–207.

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

Sendner, S., and Ruf-Oesterreicher, K. (2011). *Lambacher Schweizer 10 – Mathematik für Gymnasien Bayern: Lösungen und Materialien [Lambacher Schweizer 10 – Mathematics for the Bavarian academic school track: Solutions and materials]*. Stuttgart: Ernst Klett Verlag.

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., Kostovičová, L., and Vallée-Tourangeau, F. (2015). 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

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., and Gage, J. (2015). What can education learn from real-world communication of risk and uncertainty? *Math. Enthus.* 12, 4–10.

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

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

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

Weber, P. (2016). *Natürliche Häufigkeiten – Chancen und Grenzen aus fachwissenschaftlicher und Fachdidaktischer Sicht [Natural Frequencies – Benefits and Limits From a Mathematical and an Educational Perspective]*. Master's thesis, University of Regensburg.

Wegwarth, O., and Gigerenzer, G. (2013). Overdiagnosis and overtreatment: evaluation of what physicians tell their patients about screening harms. *JAMA Intern. Med.* 173, 2086–2088. doi: 10.1001/jamainternmed.2013.10363

Woike, J. K., Hoffrage, U., and Martignon, L. (2017). Integrating and testing natural frequencies, naïve Bayes, and fast-and-frugal trees. *Decision* 4, 234–260. doi: 10.1037/dec0000086

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: frequency or nested sets? *Exp. Psychol.* 50, 97–106. doi: 10.1027//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: 20 February 2019
doi: 10.3389/fpsyg.2019.00267

How to Improve Performance in Bayesian Inference Tasks: A Comparison of Five Visualizations

Katharina Böcherer-Linder[1] *and Andreas Eichler*[2]*

[1]Institute of Mathematics, University of Freiburg, Freiburg, Germany, [2]Institute of Mathematics, University of Kassel, Kassel, Germany

Bayes' formula is a fundamental statistical method for inference judgments in uncertain situations used by both laymen and professionals. However, since people often fail in situations where Bayes' formula can be applied, how to improve their performance in Bayesian situations is a crucial question. We based our research on a widely accepted beneficial strategy in Bayesian situations, representing the statistical information in the form of natural frequencies. In addition to this numerical format, we used five visualizations: a 2 × 2-table, a unit square, an icon array, a tree diagram, and a double-tree diagram. In an experiment with 688 undergraduate students, we empirically investigated the effectiveness of three graphical properties of visualizations: area-proportionality, use of discrete and countable statistical entities, and graphical transparency of the nested-sets structure. We found no additional beneficial effect of area proportionality. In contrast, the representation of discrete objects seems to be beneficial. Furthermore, our results show a strong facilitating effect of making the nested-sets structure of a Bayesian situation graphically transparent. Our results contribute to answering the questions of how and why a visualization could facilitate judgment and decision making in situations of uncertainty.

Keywords: epistemic uncertainty, Bayesian situations, judgment and decision making, visualization of statistical information, nested-sets structure

OPEN ACCESS

Edited by:
Gorka Navarrete,
Adolfo Ibáñez University, Chile

Reviewed by:
Lenka Kostovicova,
Slovak Academy of Sciences (SAS), Slovakia
Sebastian Hafenbrädl,
University of Navarra, Spain

***Correspondence:**
Andreas Eichler
eichler@mathematik.uni-kassel.de

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

Received: *16 March 2018*
Accepted: *28 January 2019*
Published: *20 February 2019*

Citation:
Böcherer-Linder K and Eichler A (2019) How to Improve Performance in Bayesian Inference Tasks: A Comparison of Five Visualizations. Front. Psychol. 10:267. doi: 10.3389/fpsyg.2019.00267

INTRODUCTION

A typical case of judgment and decision making in a situation of epistemic uncertainty emerges when a medical diagnosis test yields a positive result. In this situation, the physician has to make a judgment and a decision about the health status of his or her patient and possibly about further treatment. An often cited example is shown in **Figure 1** (cf. Johnson and Tubau, 2015, p. 3).

The uncertainty of the given situation is twofold. On one hand, medical diagnosis tests comprise an aleatory uncertainty, that is, an uncertainty that could not be changed in a given situation, similar to the probability distribution of dice. On the other hand, the uncertainty that a physician has concerning the health status of a patient represents epistemic uncertainty that is based on lack of knowledge (for both types of uncertainty, cf. Baraldi et al., 2014). Epistemic uncertainties can be changed by further information, such as a positive result on a diagnosis test, since the test result changes the physician's knowledge status. For this reason, appropriately processing important information in a situation, such

"10% of women at age forty who participate in a study have a particular disease. 60% of women with the disease will have a positive reaction to a test. 20% of women without the disease will also test postitive."

Calculate the probability of having the particular disease given a positive test result.

FIGURE 1 | A medical context of judgment and decision making under uncertainty.

"10 out of 100 women at age forty who participate in a study have a particular disease. 6 of the 10 women with the disease will have a positive reaction to a test. 18 of the 90 women without the disease will also test positive."

Calculate the proportion of infected people among those testing positive.

FIGURE 2 | The medical context of **Figure 1** with statistical information represented by natural frequencies.

as a medical diagnosis test, is a crucial competence of professionals as well as laymen when confronted with epistemic uncertainty (e.g., Koller and Hoffrage, 2015). Similar judgments and decisions are also essential for lawyers, if evidence is given concerning a person being guilty or innocent (e.g., Satake and Murray, 2014), as well as in other professions (Hoffrage et al., 2015; Mellers et al., 2017). By contrast, a failure of processing information in a situation of epistemic uncertainty can lead to misjudgments and severe consequences (e.g., Stine, 1998; Schneps and Colmez, 2013). For this reason, the main aim of this paper is to contribute to answering the question of how to facilitate judgment and decision making in situations of epistemic uncertainty.

A main model to process information in a situation of epistemic uncertainty is Bayes' formula, as shown in **Figure 1**. This formula allows a quantitative judgment for one of the several possible hypotheses. Therefore, we call a situation as shown in **Figure 1** a "Bayesian situation". In our example, there are two possible hypotheses: having the disease or not. If H is the hypothesis and D the information, the epistemic uncertainty $P(H)$ could be replaced by $P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D|H) \cdot P(H) + P(D|\bar{H}) \cdot P(\bar{H})}$. Unfortunately, both professionals and laymen often fail to process information in a Bayesian situation (Eddy, 1982; Hoffrage et al., 2000; Ellis et al., 2014). Based on the importance of appropriately dealing with epistemic uncertainties in different professions (see above), it has been reported that different strategies, such as using natural frequencies and using visualization, can greatly enhance performance in these situations (Gigerenzer and Hoffrage, 1995; Brase, 2009). A meta-analysis by McDowell and Jacobs (2017) found that the natural frequencies strategy increases participants' performance from about 4% to about 24%. The representation of the situation in **Figure 1** with natural frequencies is shown in **Figure 2**.

By contrast, discussion about a facilitating effect of visualizing statistical information in Bayesian situations is more ambiguous. Actually, it is an ongoing question which kind of visualization effectively increases people's performance in Bayesian situations (e.g., Binder et al., 2015). Furthermore, it is an open question why visualizations are essential for improvement, or rather, which properties of the visualizations are essential for improvement (e.g., Brase, 2009, 2014; Sirota et al., 2014b). For this reason, the main aim of our study is to investigate properties of visualizations that could potentially increase people's performance in Bayesian situations beyond the effect of natural frequencies; thus, the study could contribute to a prescriptive theory of improving statistically driven judgment and decision making in situations of epistemic uncertainty.

In this paper, we first discuss in detail visualizations of Bayesian situations and their possible facilitating effect, based on which we propose three hypotheses. These three hypotheses were investigated in an experiment that we conducted in a sample of 688 undergraduate students.

RESEARCH QUESTIONS AND HYPOTHESES CONCERNING VISUALIZING STATISTICAL INFORMATION

In former research, three graphical properties of visualizations were considered to be beneficial in the context of boosting performance in Bayesian situations. The three main ideas are representing the statistical information area proportionally, using discrete objects, and making the nested-sets structure of a Bayesian situation transparent. In the following, we discuss these three approaches of boosting performance in Bayesian reasoning tasks. Performance was measured by the ability to solve Bayesian reasoning tasks in the frequency format (**Figure 2**). Based on our former research results (Böcherer-Linder and Eichler, 2017), this discussion leads to a research question and a related hypothesis for each of the three approaches.

Starting with the 2 × 2-table that contains the information in terms stated as simply as possible, a unit square combines the properties of the 2 × 2-table with an area-proportionality (see **Figures 3A,B**). Second, the icon array combines the properties

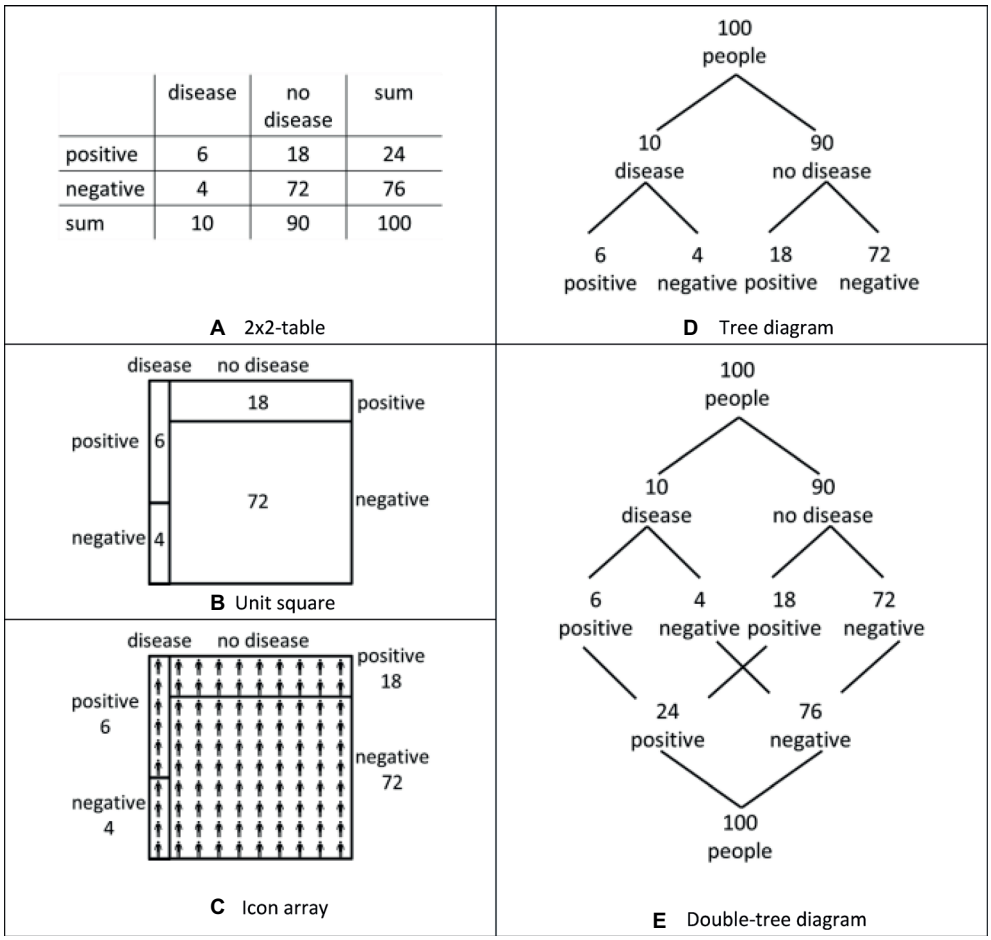


FIGURE 3 | Five visualizations of the medical diagnosis situation, that is, a 2×2-table **(A)**, a unit square **(B)**, an icon array **(C)**, a tree diagram **(D)**, and a double-tree diagram **(E)**.

of the unit square showing discrete objects (**Figure 3C**). Finally, the double-tree diagram combines the properties of the tree diagram and the property of a graphical transparency of nested sets (**Figures 3D,E**).

Area-Proportionality (Comparison of the 2 × 2-Table and the Unit Square)

One idea about beneficial graphical properties of visualization refers to representing the statistical information area proportionally (e.g., Tsai et al., 2011; Micallef et al., 2012): "this accurate, proportional representation is considered a key feature of what makes a good visual aid" (Talboy and Schneider, 2017, p. 375). Theoretical arguments for area-proportional visualizations, such as the unit square (**Figure 3B**), are often formulated based on mathematical considerations: "Rectangular areas correspond to probabilities and can be used to calculate their numerical value and to determine the Bayes relation" (Oldford, 2003, p. 1). Area-proportional visualizations increased performance in Bayesian reasoning tasks in Tsai et al. (2011), whereas area-proportionality did not prove to be a facilitating factor in Micallef et al. (2012) or Talboy and Schneider (2017).

We refer to the unit square (see **Figure 3B**), which was an effective visualization in our former research (Böcherer-Linder and Eichler, 2017) and which is an area-proportional visualization. In contrast to the unit square, the 2 × 2-table (see **Figure 3A**) is a visualization of the same graphical style (cf. Khan et al., 2015) without the property of area-proportionality. Therefore, by comparing the unit square with the 2 × 2-table, the first research question is whether the area-proportionality of the unit-square has an effect on performance in Bayesian reasoning tasks. Following the view of Talboy and Schneider (2017) and Oldford (2003), we thus hypothesize the following:

Hypothesis 1: A unit square will be more effective than a 2 × 2-table with respect to performance in Bayesian reasoning tasks.

Discrete Objects (Comparison of the Unit Square and the Icon Array)

A second idea about beneficial graphical properties of visualizations refers to using representations of "real, discrete and countable" objects (Cosmides and Tooby, 1996, p. 33). This graphical property has been claimed to be helpful because

it imitates the natural sampling situation and "help tap into the frequency coding mechanisms of the mind" (Brase, 2009, p. 369). The theoretical background of this approach is the ecological rationality account assuming that people perform better if the problem presentation resembles a real environmental situation (Gigerenzer, 2017). Realizations of visualizations with real, discrete, and countable objects include icon arrays (see **Figure 3C** for an example) and frequency grids. Beneficial effects have been observed for these kinds of visualizations, for example, in Sedlmeier and Gigerenzer (2001), Brase (2009, 2014), and Garcia-Retamero and Hoffrage (2013), but not in Sirota et al. (2014b, Experiment 1). Concerning iconicity, that is, the extent to which the icons resemble the represented objects, Sirota et al. (2014b) and Brase (2014) did not find any positive effect.

We refer again to the unit square that proved to be an effective visualization (Böcherer-Linder and Eichler, 2017). Following the idea of representing discrete objects to boost performance, we pose the research question of whether the beneficial effect of the unit square can further be enhanced by adding discrete objects into the fields, which led us to the design of the icon array as shown in **Figure 3C**. According to the theoretical considerations above, we expect a beneficial effect because of the additional discrete objects in the visualization:

Hypothesis 2: An icon array will be more effective than a unit square with respect to performance in Bayesian reasoning tasks.

Graphical Transparency of Nested Sets (Comparison of the Tree Diagram and Double Tree Diagram)

A third idea about beneficial graphical properties of visualization refers to the transparent representation of the nested-sets structure of Bayesian situations. The theoretical background for this approach is the nested-sets account claiming that "any manipulation that increases the transparency of the nested-sets relation should increase correct responding" (Sloman et al., 2003, p. 302). Examples of graphical representations of a nested-sets structure or rather nested-sets relation include Euler diagrams (e.g., Sloman et al., 2003, p. 298), roulette-wheel diagrams (e.g., Yamagishi, 2003, p. 98), and unit squares (**Figure 3B**), which are close to visualizations called treemaps (Shneiderman, 1992) or identical to visualizations called mosaic displays (Friendly, 2002) or eikosograms (Oldford, 2003). Beneficial effects have been observed for these kinds of visualizations, for example, in Sloman et al. (2003) and Yamagishi (2003) but not in Brase (2009, 2014).

In our own research (Böcherer-Linder and Eichler, 2017), we argued that the tree diagram shows a weak graphical transparency of nested sets. The main argument for this assertion was that in the tree diagram, subset relations are generally visualized by connecting branches but that no branch exists connecting the subset and the set that are necessary to apply Bayes' formula, for example, the subset "infected and tested positive" and the set "all tested positive" (see **Figure 3D**). As a consequence, performance was not as high as for a diagram with transparent nested sets (Böcherer-Linder and Eichler, 2017).

Following Khan et al. (2015), the tree diagram and the unit square that we compared in Böcherer-Linder and Eichler (2017) represent different styles of visualizations, that is, a *Branch style* and a *Nested style*. However, a tree diagram can also simply be transformed into a visualization with high transparency of nested-sets structure by adding the missing branches: in a double-tree diagram (**Figure 3E**), the set or node "infected" (24) is indeed connected with the subset or node "infected among those testing positive" (6). For this reason, it is possible to compare two visualizations of the same style that differs mostly concerning the transparency of the nested sets. Thus, the question arises whether a double-tree diagram is indeed more effective because of its graphical transparency of nested sets. Following the prediction of the nested-sets account (see above), we formulate our third hypothesis:

Hypothesis 3: A double-tree diagram will be more effective than a tree diagram with respect to performance in Bayesian reasoning tasks.

For our design, the following two comments are noteworthy: First, each of the five visualizations used in this study (**Figure 3**) showed beneficial effects in former research compared to no visualization at all (e.g., 2 × 2-table: Binder et al., 2015; Talboy and Schneider, 2017; unit square: Tsai et al., 2011; Talboy and Schneider, 2017; icon array: Brase, 2009, 2014; tree diagram: Sedlmeier and Gigerenzer, 2001; Binder et al., 2015; double-tree diagram: Wassner, 2004). Thus, each visualization we used is an effective visualization when used in that context. Second, the five visualizations show a missing "numerical equivalence" among the visualizations that could potentially impact the results. We discuss this aspect in the method section.

By testing these three hypotheses, we want to push forward the question of whether adding or reducing one of the three properties—"area-proportionality," "discrete objects," and "transparent nested sets"—to a specific visualization enhances or impedes performance. The selection of two visualizations for each of the three comparisons was based on the consideration of comparing visualizations that are graphically as similar as possible but differ in the property under consideration. The results of the study may shed light on the question of whether the design of a specific visualization could be further enhanced by referring to the graphical properties of "area-proportionality," "discrete objects," and "transparency of nested sets." By this, we seek to contribute to the question of how judgment and decision-making processes might be improved in situations of epistemic uncertainty.

EXPERIMENT

Method

Participants: The participants were 688 undergraduates at the University of Kassel (Germany) and were enrolled in a course of mathematics education for primary schools. This course does not include the five visualizations (**Figure 3**) and the Bayes' rule in the curriculum. We determined the approximate

number of participants by *a priori* power analysis (G*Power) referring to a t-test (one-tailed) for our three directional hypotheses ($\alpha < 0.05$, $\beta > 0.8$, and Cohen's $d < 0.3$). The participants were randomly assigned to a 2 × 2-table ($N = 147$), a unit square ($N = 150$), an icon array ($N = 146$), a tree diagram ($N = 125$), and a double-tree diagram ($N = 120$). We had no control group, since former research showed the effectiveness of each of the five visualizations compared to no visualization (see "Experiment"). We collected the participants' data in two waves. We checked if there were differences between the two samples but did not find significant results.

Materials and procedure: To investigate the effectiveness of the five visualizations, we used five tests where the tasks, context stories, and the presented statistical data were the same, and only the visualizations differed. In **Figure 4**, we show the wording of the test items and the visualizations. Note that the original test items only showed one of the visualizations in each case. We chose a task format where we asked participants to calculate proportions and to write the solution as a fraction. Thus, we focused on applying Bayes' formula and not on the interpretation of probability. Additionally, we decided not to give the natural frequencies in the text (except the total sample size) but only within the visualizations. Therefore, problems could only be solved by reading the information from the visualizations.

To introduce the visualizations, we gave a brief description of the visualizations on the front pages of the questionnaires. This description only explained how to read out simple information from the visualizations but not how to solve Bayesian reasoning problems. For the tree diagram and unit square, these descriptions were identical to Böcherer-Linder and Eichler (2017, p. 5) and were analogous for the double-tree diagram, 2 × 2-table, and icon array.

We designed these five visualizations (**Figures 3**, **4**) with the idea that each of the visualizations has its own characteristics and we did not focus on numerical equivalence; for example, the 2 × 2-table represents the sums that are not shown in the unit square, and the double-tree diagram naturally bears more numerical information than the tree diagram. Since the participants had to read out the information from the visualizations, we additionally controlled whether each visualization was suitable for reading out simple information or sums over represented summands. Indeed, nearly every participant could answer questions such as, "How many people are not infected?" Additionally, the unit squares with column-sums (in Böcherer-Linder and Eichler, 2017) and without column-sums (in this research) showed similar effects. Therefore, we could exclude effects of the introductory description or of more or less numerical information within the visualizations.

In the icon arrays (**Figure 4**), we used icons with different degrees of iconicity (cf. Sirota et al., 2014b). The icon arrays of the items "Medical diagnosis test" and "Snowdrops" represent icons of low iconicity, and the icons of the items "Flowers" and "Clothes" represent icons of high iconicity. Since no effect

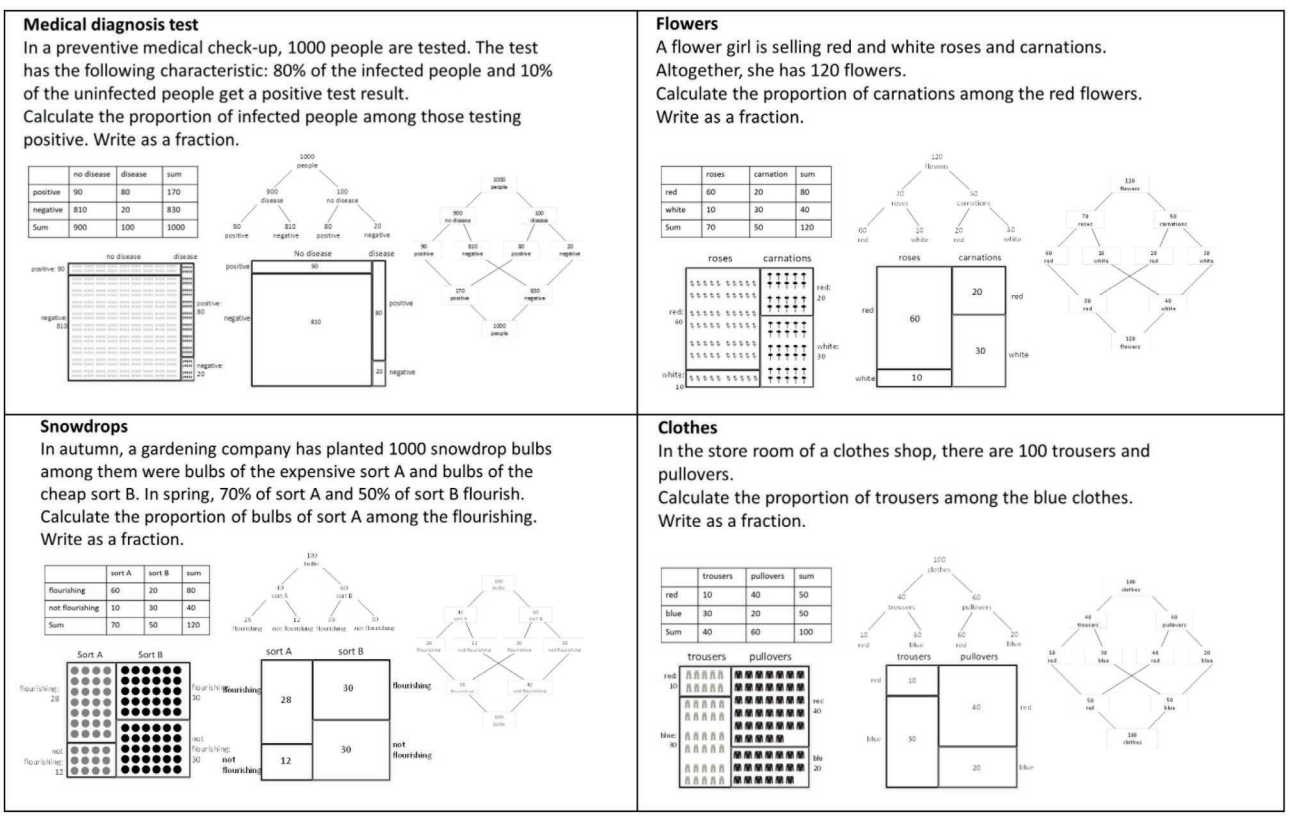


FIGURE 4 | The test-items for investigating students' performance when solving Bayesian reasoning tasks. The original test-items showed only one visualization.

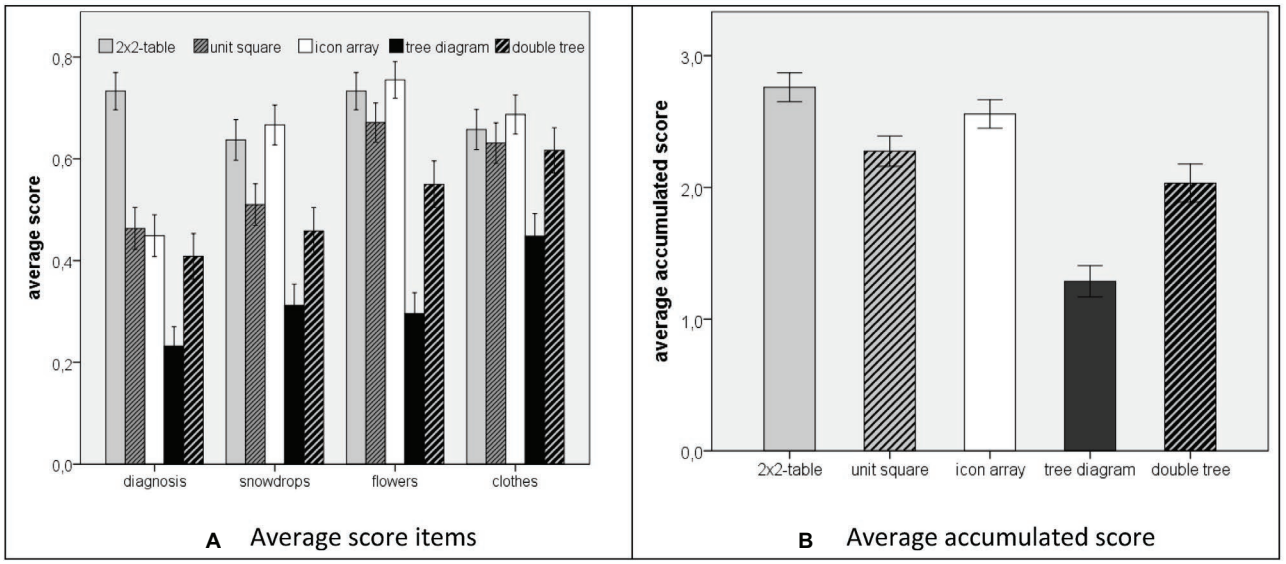


FIGURE 5 | Participants performance when solving Bayesian reasoning tasks **(A)** score for single items; **(B)** accumulated score). The error bars indicate one standard error of the mean.

of iconicity was observed in former research (Brase, 2014; Sirota et al., 2014b), we also did not expect any effect.

Since the participants were sitting close to each other when working on the tests, we arranged the items in different orders to avoid participants being influenced by each other. In our former research, where we used similar items (Böcherer-Linder and Eichler, 2017), we did not observe any effect from the order of the items. In this experiment, the seating arrangement of the participants and the order of the items was: unit square (flowers, diagnosis, clothes, snowdrops)–tree diagram (diagnosis, clothes, flowers, snowdrops)–icon array (flowers, diagnosis, clothes, snowdrops)–double tree (diagnosis, clothes, flowers, snowdrops)–2 × 2-table (snowdrops, flowers, diagnosis, clothes).

The experiment was carried out in accordance with the University Research Ethics Standards. Participation was voluntary without financial incentives, and anonymity was guaranteed. The data of our study are available at the Open Science Framework[1].

Results

An answer was rated as correct when the proportion was equal to the exact value, that is, when the fraction had a correct numerator and denominator. **Figure 5A** illustrates the results for each of the four items and **Figure 5B** shows the accumulated score for the four items. Since the four items showed an acceptable reliability for each of the five visualizations (Cronbach's Alpha: $\alpha_{2\times 2\text{-table}} = 0.685$; $\alpha_{\text{unit square}} = 0.685$; $\alpha_{\text{icon array}} = 0.658$; $\alpha_{\text{tree}} = 0.695$; $\alpha_{\text{double tree}} = 0.810$), we summarized over the four items in each case and investigated the hypotheses by comparing the accumulated scores (max = 4).

A Shapiro–Wilk test yielded a nonnormal distribution of the data. However, since the subgroups were large enough, we can assume robustness of the t-tests and of ANOVA with regard to nonnormality (Glass et al., 1972; Schmider et al., 2010). Therefore, we first conducted one-tailed t-tests to test the three hypotheses, which were directional and aimed at pairwise comparisons. Second, we applied an ANOVA for an additional exploratory analysis of our data.

Concerning *hypothesis 1*, we found a result contrary to the direction of the hypothesis. The students' performance using the 2 × 2-table was 69.0% ($M = 2.76$, $SD = 1.33$), whereas their performance using the unit square was only 56.5% ($M = 2.26$, $SD = 1.41$). Therefore, the 2 × 2-table was significantly more effective than the unit square ($t(293.619) = 3.142$, two-tailed: $p < 0.01$, Cohen's $d = 0.37$). Thus, *hypothesis 1* could not be confirmed. However, we found a significant result in the opposite direction.

Concerning *hypothesis 2*, the students' performance was higher when the information was presented in the icon array (63.9% correct solutions, $M = 2.56$, $SD = 1.31$) compared to information presentation with a unit square (56.5%; $M = 2.26$, $SD = 1.41$). The difference was significant ($t(294.238) = 1.882$, $p < 0.05$) with a small effect (Cohen's $d = 0.22$).

Concerning *hypothesis 3*, the percentage of correct solutions of students using the double tree (50.8%, $M = 2.03$, $SD = 1.58$) was significantly higher than the tree diagram (32.2%; $M = 1.28$, $SD = 1.34$, $t(232.696) = 3.989$, $p < 0.001$; Cohen's $d = 0.51$). Thus, *hypothesis 3* was confirmed.

Since we administered the five visualizations in one sample of students, we further analyzed the relations between the performances in the five conditions in an exploratory way using an ANOVA that yielded a significant result, $F(4) = 22.42$, $p < 0.001$. Post-hoc t-tests (two-tailed) with a Bonferroni correction were significant concerning a comparison of the tree diagram with each of the other four diagrams, that is, with the double-tree diagram ($M = 2.03$; $SD = 1.58$, $t(232.996) = 3.989$; $p^{\star} = 10p < 0.001$

[1]https://osf.io/g2wx7/

regarding the Bonferroni correction when testing 10 differences between two visualizations), with the 2 × 2-table (M = 2.76; SD = 1.33; $t(261.881) = 9.980$; $p^\star < 0.001$), the unit square (M = 2.26; SD = 1.41; $t(268.553) = 5.824$; $p^\star < 0.001$), and the icon array (M = 2.56; SD = 1.32; $t(261.547) = 7.878$; $p^\star < 0.001$). The effect sizes (Cohen's d) of the differences in participants' performance to solve Bayesian reasoning tasks between the tree diagram and the other diagrams were mostly high ($d_{\text{tree/double-tree}} = 0.51$; $d_{\text{tree/2 × 2-table}} = 1.25$; $d_{\text{tree/unit-square}} = 0.82$; $d_{\text{tree/icon-array}} = 1.08$), which indicated that students' performance was considerably lower when the information was presented in the tree diagram compared to each of the other four visualizations.

Post-hoc t-tests were significant concerning a comparison of the double tree diagram (M = 2.03; SD = 1.58) with the 2 × 2-table (M = 2.76; SD = 1.33; $t(232.385) = 4.078$; $p^\star < 0.001$) and the icon array (M = 2.56; SD = 1.32; $t(231.089) = 2.959$; $p^\star < 0.05$). These results indicated that students' performance was lower when the information was presented in the double tree diagram compared to the 2 × 2-table and the icon array.

DISCUSSION

There is some research evidence that supports the claim that visualizations can have an additional beneficial effect on dealing with Bayesian situations beyond representing statistical information by natural frequencies (Garcia-Retamero and Hoffrage, 2013; McDowell and Jacobs, 2017). Since research results are ambiguous with regard to which specific properties of a visualization are beneficial, we investigated five visualizations that vary concerning their style (Khan et al., 2015), form, and, particularly, concerning three properties that were found to be potentially facilitating when dealing with Bayesian situations, that is, the area-proportional representation of the statistical information, the display of discrete and countable entities, and the graphical transparency of nested sets.

First, the comparison of the 2 × 2-table and the unit square yielded no additional beneficial effect of area proportionality. To the contrary, our results imply that the 2 × 2-table had a significant positive effect compared to the unit square with a small effect (Cohen's $d = 0.37$). This result is interesting since area proportionality is the main graphical difference between the unit square and the 2 × 2-table, whereas both visualizations make the nested-sets structure transparent and show no countable objects. Our results are different from Micallef et al. (2012) who found no difference in performance between visualizations that were partly area-proportional and partly not. Additionally, in an intervention study, Talboy and Schneider (2017, p. 379), who compared the 2 × 2-table and a unit square in a training study, found that "those who were trained with graphs […] performed comparably overall with those who were trained with tables." One explanation for the unexpected result in this study is the different degree of familiarity of the 2 × 2-table and the unit square. In German schools, the 2 × 2-table is a familiar visualization. Accordingly, 86% of the participants indicated knowing the 2 × 2-table. In contrast, only 34% of the participants indicated knowing a visualization like the unit square. An alternative explanation for the unexpected supremacy of the 2 × 2-table could be supposed concerning the context. Thus, the difference between the 2 × 2-table and the unit square seems to be influenced by the item "diagnosis" that had the most extreme distribution of the data in the 2 × 2-situation compared to the other three items. However, both assumptions need to be investigated in further research.

Second, the icon array outperformed the unit square. This result seems to be a consequence of visualizing countable and discrete entities (icons), since other potentially effective properties remained constant (area-proportionality and transparency of nested sets). This result is in accordance with Brase (2009) who found a positive effect when adding dots into Euler diagrams. However, taking into account the small effect (Cohen's $d = 0.22$), we would not go as far as Brase (2014) in claiming that icon representations "are the most powerful pictorial technique currently known for facilitating correct Bayesian reasoning" (p. 93), since in our study, the effect of the graphical transparency of nested sets was higher than the effect of representing discrete objects.

Third, the double tree diagram outperformed the tree diagram. An explanation of this result is that the double tree diagram makes the nested-sets structure of a Bayesian situation graphically transparent in contrast to the tree diagram. The effect of making the nested-sets structure transparent was prominent in our results. Furthermore, the unit-square, the icon array, and the 2 × 2-table make the nested-sets structure graphically transparent by neighboring fields that have to be considered in a Bayesian situation (cf. Böcherer-Linder and Eichler, 2017). Accordingly, the ANOVA and post-hoc t-test showed two things. First, the difference between the performances in Bayesian reasoning was high with large effects, when a visualization made the nested-sets structure of a Bayesian situation transparent (except for the comparison of the tree diagram with the double tree diagram). Although the representation of discrete objects also yielded a positive effect, our results imply that the most powerful visualization of Bayesian situations is a visualization with natural frequencies that make the nested-sets structure of a Bayesian situation graphically and numerically transparent. The effect of making the nested-sets structure transparent was constant across visualizations representing different styles identified by Khan et al. (2015) and was further constant between two visualizations of the same style (tree diagram and double-tree diagram). For this reason, the beneficial effect of making the nested-sets structure in a Bayesian situation transparent seems to be very clear.

Properties of the sample and the tasks' characteristics could have influenced our results. First, our sample consisted of university students. Thus, the results must be interpreted with this in mind, since intellectual ability seems to have an impact on performance in Bayesian situations (e.g., Johnson and Tubau, 2015), and particularly, spatial abilities might influence the effect of visualizations (Ottley et al., 2016). However, although the students were enrolled in a mathematics education course, these students' affinity to mathematics was (on average) not high, since every primary teacher in Germany has to take courses in mathematics independent of ability or motivation to learn mathematics. Second, the context of the Bayesian situations and the wording of the tasks could have affected performance (e.g., Siegrist and Keller,

2011; Böcherer-Linder et al., 2018). Actually, the different contexts in the Bayesian situations that we used influenced performance. However, concerning the focus of this paper, that is, the beneficial effect of visualizations' properties, in the Bayesian situations that we used, no interaction effect between visualization and context was found. Finally, the degree of familiarity seems to be a property of a given visualization that has to be taken into account. Further research could take differences in the mentioned properties of visualizations into account, which might also provide explanations for the significant differences between the double tree diagram and both the 2 × 2-table and the icon array.

The results that we present in this paper are based on the first beneficial strategy of representing statistical information as natural frequencies. In addition to the numerical transparency of nested sets, the second beneficial strategy refers to making the nested-sets structure of a Bayesian situation graphically transparent. The connection of both strategies resulted in a performance of about 60% in different Bayesian reasoning tasks. This is a considerable facilitating effect compared to the low performance of about 5% if the statistical information is represented by probabilities and without visual aids (cf. McDowell and Jacobs, 2017). Is this successful enough? Since dealing with Bayesian situations inappropriately, as a specific class of situations of epistemic uncertainty, could include severe consequences, for example, making inappropriate judgments and decisions in a medical diagnosis test or a jury verdict, a performance of about 60% should be increased further. For this reason, a further focus could be placed on investigating interventions that potentially further increase performance in Bayesian situations (Sirota et al., 2014a). For training studies, the question arises of which property of a visualization would yield short-run or even long-run success in dealing with Bayesian situations (cf. Sedlmeier and Gigerenzer, 2001).

Moreover, although our results imply that the icon array and the 2 × 2-table are more effective than the unit square and, particularly, more effective than the double tree diagram when performance in Bayesian situations is considered, further properties of the visualizations could be taken into account. For example, graphical differences might play a role when applying these visualizations in training. For example, if statistical information is given in text format and has to be visualized actively, the icon array, even with dots, is difficult to construct if a sample in a Bayesian situation is big and comprises, for example, 1,000 statistical entities. Furthermore, recent research has argued for a "distinction between Bayesian performance and Bayesian reasoning" (Vallée-Tourangeau et al., 2015, p. 3). In this sense, the ability to adequately judge the influence of a parameter change in a Bayesian situation (e.g., the base rate, sensitivity, and specificity in a medical diagnosis situation) could be understood as part of Bayesian reasoning (Böcherer-Linder et al., 2017). Therefore, it is an interesting question if a visualization's property is beneficial beyond performance in Bayesian situations. We hypothesize that area-proportional visualizations could be more effective than visualizations without area-proportionality when people were asked to judge a parameter change in a Bayesian situation.

CONCLUSION

For one graphical property of visualizations, area-proportionality, we could not observe any positive effect. However, additional icons yielded a positive, albeit smaller, effect. We finally showed that visualizations making the nested-sets structure of the Bayesian situation graphically transparent could improve performance in Bayesian reasoning tasks and, thus, the ability to deal with situations of epistemic uncertainty. Thus, based on our results, the most powerful property of a visualization of Bayesian situations was the graphical transparency of the nested sets structure in these situations. Our findings could inform the debate about beneficial graphical properties of visual representations of statistical information in Bayesian situations and could serve as an empirical foundation for designing interventions for improving judgment and decision making based on Bayesian reasoning for both professionals and laymen.

AUTHOR CONTRIBUTIONS

Both authors equally contributed to the research approach, to the realization of the research and the concept and writing of the paper.

FUNDING

This work was supported by the Ministry of Science, Research and the Arts of the State of Baden-Württemberg, Germany.

REFERENCES

Baraldi, P., Compare, M., and Zio, E. (2014). Uncertainty treatment in expert information systems for maintenance policy assessment. *Appl. Soft Comput.* 22, 297–310. doi: 10.1016/j.asoc.2014.05.024

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

Böcherer-Linder, K., and Eichler, A. (2017). The impact of visualizing nested sets. An empirical study on tree diagrams and unit squares. *Front. Psychol.* 7:2026. doi: 10.3389/fpsyg.2016.02026

Böcherer-Linder, K., Eichler, A., and Vogel, M. (2017). The impact of visualization on flexible Bayesian reasoning. *Av. Investig. Educ. Mat.* 11, 25–46.

Böcherer-Linder, K., Eichler, A., and Vogel, M. (2018). Die Formel von Bayes: Kognitionspsychologische Grundlagen und empirische Untersuchungen zur Bestimmung von Teilmenge-Grundmenge-Beziehungen (Bayes' formula: *principles of cognitive psychology and investigation of dealing with nested-sets relations). J. Mathe. Did.* 39, 127–146. doi: 10.1007/s13138-018-0128-1

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

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

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.

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

Friendly, M. (2002). A brief history of the mosaic display. *J. Comput. Graph. Stat.* 11, 89–107. doi: 10.1198/106186002317375631

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. (2017). "Ökologische Rationalität (ecological rationality)" in *Dorsch – Lexikon der Psychologie (Dorsch - psychological enzyclopedia)*. ed. M. A. Wirtz. Retrieved from https://portal.hogrefe.com/dorsch/oekologische-rationalitaet/ [15.03.2017]

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

Glass, G. V., Peckham, P. D., and Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev. Educ. Res.* 42, 237–288.

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

Khan, A., Breslav, S., Glueck, M., and Hornbæk, K. (2015). Benefits of visualization in the Mammography problem. *Int. J. Hum. Comput. St.* 83, 94–113. doi: 10.1016/j.ijhcs.2015.07.001

Koller, M., and Hoffrage, U. (2015). Societal perspectives on risk awareness and risk competence. *German Med. Sci.* 13:Doc08. doi: 10.3205/000212

McDowell, M., and Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychol. Bull.* 143, 1273–1312. doi: 10.1037/bul0000126

Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., and Tetlock, P. E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgm. Decis. Mak.* 12, 369–381.

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

Oldford, R. W. (2003). *Probability, problems, and paradoxes pictured by eikosograms*. Retrieved from http://www.stats.uwaterloo.ca/~rwoldfor/papers/venn/eikosograms/examples/paper.pdf [15.03.2017]

Ottley, A., Peck, E., Harrison, L., Afergan, D., Ziemkiewicz, C., Taylor, H., et al. (2016). Improving Bayesian reasoning: the effects of phrasing, visualization, and spatial ability. *IEEE Trans. Vis. Comput. Graph.* 22, 529–538. doi: 10.1109/TVCG.2015.2467758

Satake, E., and Murray, A. V. (2014). Teaching an application of Bayes' rule for legal decision-making: measuring the strength of evidence. *J. Stat. Educ.* 22, 1–29.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., and Bühner, M. (2010). Is It Really Robust? *Methodology* 6, 147–151. doi: 10.1027/1614-2241/a000016

Schneps, L., and Colmez, C. (2013). *Math on trial: how numbers get used and abused in the courtroom*. New York NY: Basic Books.

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

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.* 11, 92–99. doi: 10.1145/102377.115768

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. *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

Stine, G. J. (1998). *Acquired immune deficiency syndrome: biological, medical, social, and legal issues*. 3rd Edn. Upper Saddle River, NJ: Prentice Hall.

Talboy, A. N., and Schneider, S. L. (2017). Improving accuracy on Bayesian inference problems using a brief tutorial. *J. Behav. Decis. Mak.* 30, 373–388. doi: 10.1002/bdm.1949

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

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

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

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

ORIGINAL RESEARCH
published: 15 May 2019
doi: 10.3389/fpsyg.2019.00632

An Eye-Tracking Study of Statistical Reasoning With Tree Diagrams and 2 × 2 Tables

Georg Bruckmaier[1], Karin Binder[2], Stefan Krauss[2] and Han-Min Kufner[2]*

[1] Department of Secondary Education, University of Education, University of Applied Sciences and Arts Northwestern Switzerland, Windisch, Switzerland, [2] Mathematics Education, Faculty of Mathematics, University of Regensburg, Regensburg, Germany

Changing the information format from probabilities into frequencies as well as employing appropriate visualizations such as tree diagrams or 2 × 2 tables are important tools that can facilitate people's statistical reasoning. Previous studies have shown that despite their widespread use in statistical textbooks, both of those visualization types are only of restricted help when they are provided with probabilities, but that they can foster insight when presented with frequencies instead. In the present study, we attempt to replicate this effect and also examine, by the method of eye tracking, *why* probabilistic 2 × 2 tables and tree diagrams do not facilitate reasoning with regard to Bayesian inferences (i.e., determining what errors occur and whether they can be explained by scan paths), and *why* the same visualizations are of great help to an individual when they are combined with frequencies. All ten inferences of $N = 24$ participants were based solely on tree diagrams or 2 × 2 tables that presented either the famous "mammography context" or an "economics context" (without additional textual wording). We first asked participants for marginal, conjoint, and (non-inverted) conditional probabilities (or frequencies), followed by related Bayesian tasks. While solution rates were higher for natural frequency questions as compared to probability versions, eye-tracking analyses indeed yielded noticeable differences regarding eye movements between correct and incorrect solutions. For instance, heat maps (aggregated scan paths) of distinct results differed remarkably, thereby making correct and faulty strategies visible in the line of theoretical classifications. Moreover, the inherent structure of 2 × 2 tables seems to help participants avoid certain Bayesian mistakes (e.g., "Fisherian" error) while tree diagrams seem to help steer them away from others (e.g., "joint occurrence"). We will discuss resulting educational consequences at the end of the paper.

Keywords: Bayesian reasoning, eye tracking, 2 × 2 table, tree diagram, natural frequencies, probabilities

OPEN ACCESS

Edited by:
Gorka Navarrete,
Adolfo Ibáñez University, Chile

Reviewed by:
Miroslav Sirota,
University of Essex, United Kingdom
Manuele Reani,
The University of Manchester, United Kingdom

****Correspondence:***
Georg Bruckmaier
georg.bruckmaier@fhnw.ch

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

Received: *12 November 2018*
Accepted: *06 March 2019*
Published: *15 May 2019*

Citation:
Bruckmaier G, Binder K, Krauss S and Kufner H-M (2019) An Eye-Tracking Study of Statistical Reasoning With Tree Diagrams and 2 × 2 Tables. Front. Psychol. 10:632. doi: 10.3389/fpsyg.2019.00632

INTRODUCTION

It is relevant to one's understanding of statistical situations involving two binary uncertain events (e.g., being ill: yes/no; medical test: positive/negative) whether the information is presented in probabilities (e.g., "80%") or in natural frequencies (e.g., "8 out of 10"; Gigerenzer and Hoffrage, 1995). In the case of what is known as Bayesian reasoning situations, a meta-study found that the

change of probabilities in natural frequencies substantially increases performance rates (McDowell and Jacobs, 2017; see also Barbey and Sloman, 2007). In Bayesian reasoning situations concerning medical contexts, the prevalence (*a priori* probability) of a disease is usually given, as well as the sensitivity and false-alarm rate of a medical test (see section Statistical Situations Based on Two Binary Events for a detailed theoretical distinction between Bayesian and non-Bayesian reasoning situations). Furthermore, a good deal of the literature demonstrates that visualizations can also foster insight into Bayesian reasoning or in statistical thinking in general (Yamagishi, 2003; Steckelberg et al., 2004; Binder et al., 2015; see also **Figures 1**, **2**). In cognitive psychology—because of their relevance in real-world medical and legal decision-making (Hoffrage and Gigerenzer, 1998; Hoffrage et al., 2000; Fenton et al., 2016; Operskalski and Barbey, 2016)—Bayesian inferences stand firmly in the foreground of discussions about statistical reasoning.

In the field of statistics education, secondary school and university students have to assess and understand *all* probabilities concerning situations involving two binary events such as conjoint probabilities or (non-inverted) conditional probabilities (in such situations, 16 different probabilities can be considered, see section Statistical Situations Based on Two Binary Events). Thus in statistics classes taught at secondary schools or universities, a Bayesian inference is often treated as merely a (complicated) special case of conditional probability.

Regarding visualizations, in Germany but also in many other countries, tree diagrams and 2 × 2 tables are particularly widely implemented in textbooks on probability (see **Figure 1**; e.g., Eisentraut et al., 2008; Freytag et al., 2008; Schmid et al., 2008; Weber et al., 2018), most likely because both visualizations explicitly contain numbers and can be constructed easily by students based on typical problem wordings (neither of which is the case for, e.g., Euler diagrams or similar visualizations that rely on geometrical areas; see **Figure 2**; Weber et al., 2018). However, when the visualizations are equipped with probabilities (which in the classroom is most often the case), students unfortunately seem to struggle regardless of which of the two visualizations is used—especially concerning the notorious Bayesian inferences. Binder et al. (2015) could demonstrate that although German high school students are pretty much familiar with both visualizations, they cannot exploit tree diagrams or 2 × 2 tables with probabilities for respective inferences, and that the situation only changes when both visualizations are presented with frequencies (see **Figure 1**).

The study detailed in this paper attempts to replicate format effects concerning visualizations and goes one step further by investigating corresponding cognitive processes with the method of eye tracking. We expect with this method to be able to identify and describe typical (correct) solution strategies on the one hand, and on the other to explain specific errors frequently made by the participants. Thus our study investigates the intriguing question of why so many people struggle with probabilistic reasoning (including Bayesian), even when the widely prominent tree diagrams and 2 × 2 tables visualize the situation for them. What is wrong with these visualizations? And how do scan paths change when both visualizations are instead given with frequencies? Despite multiple calls for its use (Verschaffel et al., 2016; McDowell and Jacobs, 2017), the method of eye tracking has been applied only a few times thus far within the framework of statistical reasoning (Cohen and Staub, 2015; Reani et al., 2017; Lehner and Reiss, 2018), and not at all for analyzing format differences concerning both widely applied visualizations.

It has to be noted that most research in the field of cognitive psychology or statistics education—with a strong focus on the special case of Bayesian inferences, especially in cognitive psychology—is concerned with attempts to boost performance, for instance by changing the information format or presenting additional visualizations (see, e.g., the recent meta-analysis by McDowell and Jacobs, 2017), by implementing trainings (e.g., Sedlmeier and Gigerenzer, 2001; Steckelberg et al., 2004), or by theoretically explaining the benefit of certain tools (e.g., the discussion between proponents of the ecological rationality approach and the nested sets approach, Hoffrage et al., 2000; Pighin et al., 2016). With mathematics education in mind, the present research is in line with recent studies also conducted by our research group that look at the other side of the coin of statistical reasoning: when and why teaching fails. For instance, by focusing on participants who failed in Bayesian inferences *although* the information was displayed in terms of the favored frequencies, Weber et al. (2018) could demonstrate that due to a "fixed mindset," many of these students translated the given natural frequencies "back" into probabilities, with the consequence that they were not able to solve the task.

In the first theoretical section of the paper, we will show that Bayesian inferences are only a special case in situations with two binary uncertain events, and examine which other probabilities are regularly covered in teaching at secondary school and university. We will then explain why tree diagrams and 2 × 2 tables are both widely implemented worldwide in the actual teaching of statistics, and what is already known about typical errors that are made with regard to inferences based on those two visualizations. In this way, the rationale of our present approach combines the concept of natural frequencies and the focus on *Bayesian* reasoning from cognitive psychology with a consideration of all 16 probabilities and the choice to utilize tree diagrams and 2 × 2 tables in typical statistics education materials used at secondary school and university.

STATISTICAL THINKING

Statistical Situations Based on Two Binary Events

Bayesian situations usually refer to two binary uncertain events such as a state of health (being ill vs. not being ill) and a medical test result (e.g., positive vs. negative). In secondary school, and especially with younger children, the respective events might, for instance, be the gender of a child (female vs. male) and a certain personality trait (e.g., loves sports vs. does not love sports). In general, in such situations, 16 different probabilities can be theoretically considered, which we will illustrate with the case of the famous mammography context (that will also be applied later on as one of the two contexts in our empirical study). The

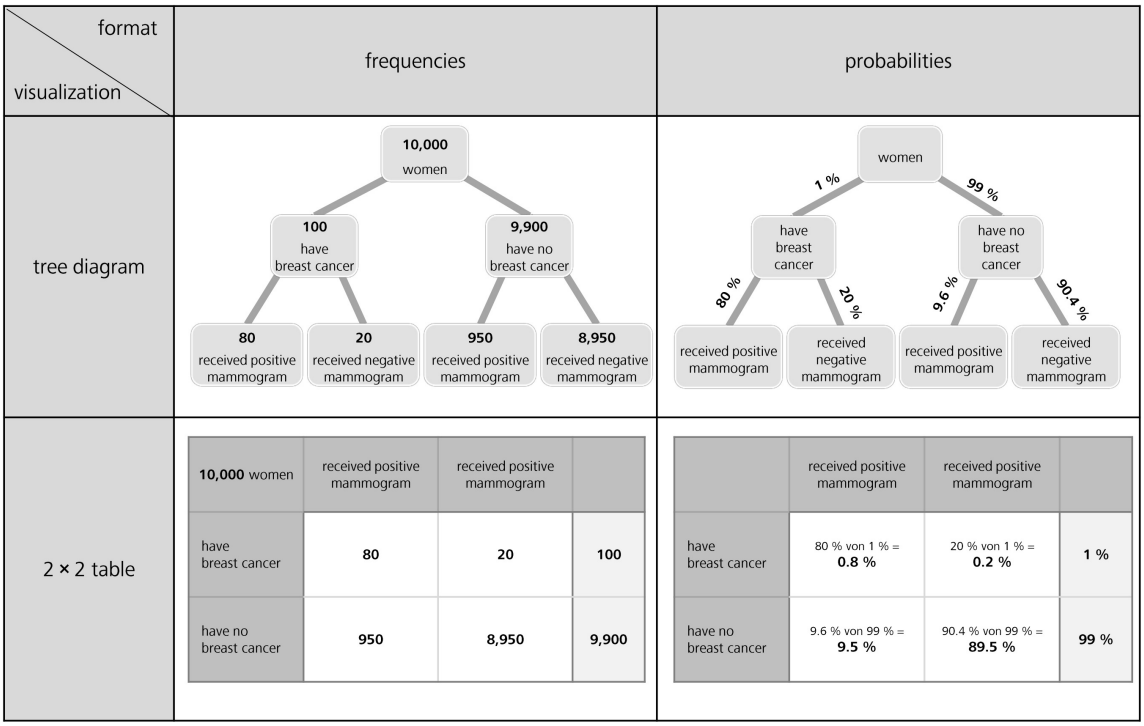


FIGURE 1 | Tree diagrams (above) and 2 × 2 tables (below), both with frequencies (left) and with probabilities (right) for the mammography context (figure adapted from Binder et al., 2015).

mammography context contains two events, each with binary values (B: having breast cancer; B: not having breast cancer; M+: positive mammogram; M−: negative mammogram), which allows for the consideration of the following probabilities:

Four probabilities taking just one event into account (marginal probabilities):
P(B), P(¬B), P(M+), P(M−),
with P(¬B) = 1 − P(B) and P(M−) = 1 − P(M+)

Four conjoint probabilities:
$P(B \cap M+)$, $P(\neg B \cap M+)$, $P(B \cap M-)$, $P(\neg B \cap M-)$

Eight conditional probabilities:
$P(M+|B)$, $P(M+|\neg B)$, $P(M-|B)$, $P(M-|\neg B)$,
$P(B|M+)$, $P(B|M-)$, $P(\neg B|M+)$, $P(\neg B|M-)$

Note that thus far, no task is given, and it is possible to describe these situations in general without the need to decide on a special inference (consequently, in the following we will strictly distinguish between the "mammography situation" *per se* and the corresponding problem/task posed). Respective inferences often require—in cognitive psychology and in the teaching of statistics as well—deducing a certain probability when at least three other probabilities are given. The most prominent examples are Bayesian inferences that involve the inversion of a given conditional probability. For instance:

Mammography problem (probability format):

The probability of breast cancer (B) is 1% for a woman of a particular age group who participates in a routine screening (P(B)). If a woman who participates in a routine screening has breast cancer, the probability P(M+|B) is 80% that she will have a positive mammogram (M+). If a woman who participates in a routine screening does not have breast cancer (B), the probability P(M+|B) is 9.6% that she will have a false-positive mammogram.

What is the probability that a woman who participates in a routine screening and has a positive mammogram has breast cancer?

The required Bayesian inference is an "inversion" in the sense that a conditional probability P(M+|B) is given and the "inverse" conditional probability P(B|M+) has to be assessed in order to "update" an *a priori* estimation [in this case P(B)]. In the light of this new evidence, Bayes' theorem yields:

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

It is well known that such solutions may be counterintuitive (especially when extreme base rates like 1% are given) and that most people (even experts like physicians) have difficulty estimating such probabilities. In the meta-analysis by McDowell and Jacobs (2017), only 4% of the participants were able to come up with correct answers concerning such inferences. However, in addition to these problematic Bayesian inversions,

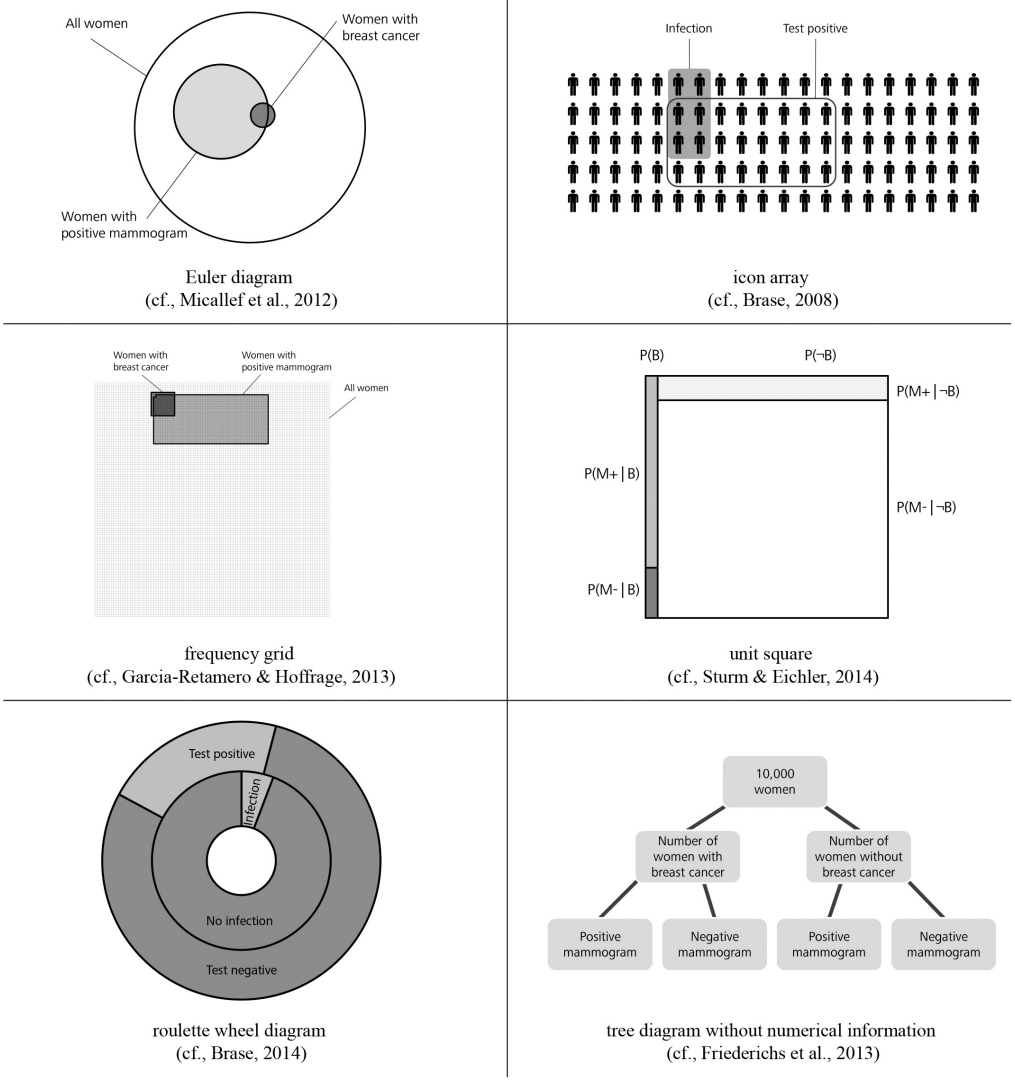


FIGURE 2 | Alternative visualizations (figure from Binder et al., 2015).

the assessment of conjoint probabilities (e.g., Fiedler, 2000) can also be difficult.

Information Formats: Probabilities vs. Frequencies

Nevertheless, situations like these can actually be taught to very young children who are not even aware of the concept of conditional probability (or probabilities in general). In German secondary schools, for instance, such situations are introduced to children as young as 10, with absolute numbers concerning a set of persons (or objects) provided, each of them having (or not having) two certain characteristics. For instance, there may be 100 students, and the two characteristics might be gender (male or female) and wearing glasses (or not). Note that when a certain sample is given, all of the 16 probabilities mentioned above can be expressed in absolute numbers that describe specific subsets. The fact that absolute numbers are much easier to grasp is exploited by the concept of *natural frequencies* (Gigerenzer and Hoffrage, 1995), which even foster insight into Bayesian inferences. Natural frequencies combine two absolute frequencies, as illustrated in the mammography problem:

Mammography problem (natural frequency format):

100 out of 10,000 women of a particular age group who participate in a routine screening have breast cancer. 80 out of 100 women who participate in a routine screening and have breast cancer will have a positive mammogram. 950 out of 9,900 women who participate in a routine screening and have no breast cancer will have a false-positive mammogram.

How many of the women who participate in a routine screening and receive positive mammograms have breast cancer?

Substantially more people are now able to find the correct solution to the problem (which is "80 out of 1,030") because

the solution becomes more obvious and the calculation is easier. In the meta-analysis by McDowell and Jacobs (2017), frequency versions of Bayesian reasoning problems can be solved on average by 24% of participants across studies and contexts. Even in more complex Bayesian problems, such as in situations involving more than one medical test or unclear test results, frequencies help people in their decision-making processes (Hoffrage et al., 2015b; Binder et al., 2018). In the last 20 years, an abundance of studies has shown the facilitating effect of frequencies for many different kinds of populations: physicians, patients, judges in court, managers, university and high school students, and even young children (Gigerenzer and Hoffrage, 1995; Hoffrage et al., 2000; Zhu and Gigerenzer, 2006; Siegrist and Keller, 2011; Hoffrage et al., 2015a; McDowell and Jacobs, 2017). Weber et al. (2018), on the other side, shed light on the question of why (on average) 76% of participants still fail even though frequencies (instead of probabilities) are provided, finding that many participants translated the given frequencies back into (more complicated) probabilities.

Natural frequencies can be obtained both by natural sampling (Kleiter, 1994) or, alternatively, by actively translating given probabilities (e.g., "80%") into expressions consisting of two absolute frequencies (e.g., "80 out of 100"). In our research—in contrast to some other scholars' work (e.g., Spiegelhalter and Gage, 2015)—we consider natural frequencies as the superordinate concept for both *empirically sampled* and *expected* frequencies. While the latter constitute frequencies that are expected in the long run (cf. Hertwig et al., 2004; Spiegelhalter and Gage, 2015; case 2 in Woike et al., 2017), empirically sampled frequencies are derived from a natural sampling process (cf. Kleiter, 1994; Fiedler et al., 2000; cases 1 and 3 in Woike et al., 2017). Whereas empirically sampled frequencies can obviously deviate from the expected ones (but are still natural frequencies), expected frequencies fit perfectly into the teaching context (here, natural frequencies usually stem from imagining a specific sample).

Furthermore, it is not only natural frequencies of Bayesian tasks that can be considered natural frequencies. Of course, on the one hand it is possible to sample all of the 16 probabilities mentioned above in terms of natural frequencies (by natural sampling). And, on the other hand, if probabilities are given, all of them can actively be translated into natural frequencies as a didactical tool (by researchers, teachers, or clever students, who realize that only an arbitrary sample functioning as reference set has to be imagined first).

Number-Based Visualizations: 2 × 2 Tables and Tree Diagrams

In their research articles, scholars often use 2 × 2 tables (Goodie and Fantino, 1996; Dougherty et al., 1999; Fiedler et al., 2000) or tree diagrams (Kleiter, 1994; Gigerenzer and Hoffrage, 1995; Mandel, 2014; Navarrete et al., 2014) to illustrate Bayesian reasoning situations to their peers. Both visualizations are also very prominent in the context of statistical education at secondary school and university. Interestingly, the effects of these visualizations on participants' performance have only rarely been tested empirically thus far (for a discussion, e.g., see Binder et al., 2015). With the numbers from the mammography context above, there are generally four possible different visualizations of this kind (see **Figure 1**). The cause for the calculations in the cell at the below right is explained in issue 1 (see later in section Number-Based Visualizations: 2 × 2 Tables and Tree Diagrams).

Why are these visualizations so prominent, especially in the context of teaching? Note that in contrast to most other visual aids (see **Figure 2**), 2 × 2 tables and tree diagrams usually explicitly contain numerical information and, furthermore, both can be equipped with frequencies or with probabilities (**Figure 1**). The decisive advantage for teaching and learning, however, is that teachers and students can easily construct all of these visualizations themselves. Note that "non-numerical" visualizations such as Euler diagrams (e.g., Sloman et al., 2003; Brase, 2008; Micallef et al., 2012; Sirota et al., 2014b), roulette wheel diagrams (e.g., Yamagishi, 2003; Brase, 2014), or unit squares (Böcherer-Linder and Eichler, 2017), all of which are based on geometrical areas (**Figure 2**), require a substantial effort to be produced (i.e., sometimes the size of the specific areas needed for the visualizations can only be calculated when the task is already solved). Furthermore, it is not always convenient to display extreme base rates by a geometrical area. For instance, in a true-to-scale unit square, the prevalence of 1% would no longer be visible. Along the same lines, for displaying the mammography problem with an icon array (Brase, 2008, 2014; Sirota et al., 2014b; Zikmund-Fisher et al., 2014), which is based on small symbols instead of geometrical areas, the student (or teacher) would have to draw 10,000 icons.

It is important to note that, in principle, all visualizations appearing in **Figures 1**, **2** allow for the assessment of *all* of the 16 probabilities above (which is also true for all typical, purely textual formulations of Bayesian tasks). Furthermore, one can present not only "normal" tree diagrams or 2 × 2 tables, but also ones with highlighted branches or nodes (see Binder et al., 2018) or cells. Cognitive load theory (Sweller, 2003) would suggest that according to the signaling principle, highlighting the relevant branches, nodes, or cells might improve performance of participants (Mautone and Mayer, 2001; Mayer, 2008). Furthermore, a combination of textual and visual information could shed more light on the redundancy principle of multiple information sources, which is addressed in the cognitive load theory and the cognitive theory of multimedia learning (Mayer, 2005). The redundancy principle says, in short, that the elimination of any redundant information may enhance learning (see Sweller, 2003; Mayer, 2005) because of a reduction of the extraneous cognitive load (also see Discussion).

Concerning the four visualizations of **Figure 1** that are widely used in teaching and that we will also implement in our empirical study (for the final stimuli, see **Figure 4**), some theoretical details have to be clarified:

(1) *2 × 2 tables cannot present conditional probabilities (only tree diagrams can)*:
Concerning the probability format, it is obvious that the probabilities provided in a Bayesian task *cannot* be placed directly into a 2 × 2 table, since 2 × 2 tables contain

conjoint probabilities but not conditional ones. Therefore, while the conditional probabilities given in a Bayesian task can be placed directly on the branches of a tree diagram, 2 × 2 tables principally display different pieces of information (see **Figure 1**).

This feature of 2 × 2 tables makes them simpler (compared to tree diagrams) in terms of the calculations to be performed, at least for Bayesian inferences based on probabilities, because a part of the calculation has already been performed in order to complete the 2 × 2 table (as indicated in small letters in **Figure 1** in the cell below right). Note that only a tree diagram with probabilities requires Bayesian calculations according to formula (1), while in 2 × 2 tables the following calculation is sufficient for the resulting conditional probabilities:

$$P(B|M+) = \frac{P(B \cap M+)}{P(M+)} = \frac{0.8\%}{0.8\% + 9.5\%} \approx 7.8\% \quad (2a)$$

Consequently, since Bayesian inferences imply the aspect of inversion, it is interesting to consider whether inferences based on 2 × 2 tables containing probabilities can be called "Bayesian" at all (e.g., Binder et al., 2015, but see the short menu in Gigerenzer and Hoffrage, 1995). Therefore, in our experiments only one marginal distribution is shown (see **Figure 4**) because displaying the other one in addition would allow simply to dividing the numbers in two cells for all conditional probabilities. Thus, inverted and non-inverted conditional probabilities could not be distinguished any longer.

(2) *Concerning 2 × 2 tables, scan paths (gaze behavior) should not depend on information format*:
Concerning possible scan paths, it is important to note that, regarding 2 × 2 tables (see below in **Figure 1**), exactly the same cells would have to be inspected in both formats for all 16 possible inferences. In contrast, probabilities in tree diagrams are depicted at the branches and absolute frequencies in the nodes, thus requiring slightly deviating scan paths in the two formats. For the 2 × 2 table presented with frequencies of the mammography context, similar to formula (2a), two frequencies (instead of probabilities) have to be added to obtain the denominator in formula (2b):

$$P(B|M+) = \frac{\#(B \cap M+)}{\#(M+)} = \frac{80}{80 + 950} \approx 7.8\% \quad (2b)$$

(3) *Frequentistic visualizations are more flexible than textual natural frequency versions*:
Notably, both frequentistic visualizations (see left side in **Figure 1**) contain absolute frequencies, implying that natural frequencies of the type "x out of y" (i.e., natural frequencies always consist of two absolute frequencies) would have to be combined by first relating two absolute numbers (x and y) in any case. However, this necessity makes frequency visualizations flexible, since the absolute frequencies displayed in **Figure 1** can be combined to multiple kinds of natural frequencies (e.g., "80 out of 100," "100 out of 10,000," "80 out of 10,000").

(4) *2 × 2 tables and tree diagrams display more statistical information than textual wording*:
Furthermore, it is striking that in all four visualizations (**Figure 1**), *more* numerical information is displayed than in the corresponding mammography wordings (specifically, statistical information on the respective counter events is included). However, concerning Bayesian inferences, this additional information can usually be disregarded.

(5) *Non-inverted vs. inverted (Bayesian) conditional probabilities*:
Most importantly, with respect to Bayesian reasoning, tree diagrams (above in **Figure 1**) entail a specific order of subsetting: First, the sample is divided according to state of health, then according to test result (an inverse tree diagram can easily be imagined by first dividing the sample according to M+ and M−, and subsequently according to the state of health). In order to mirror this structure in the corresponding 2 × 2 tables, we deliberately presented only one of the two marginal distributions (in both formats, see **Figure 4**). As a consequence, we can distinguish in all four visualizations between "normal" conditional probabilities and inverse conditional probabilities in the following way: Non-inverted conditional probabilities (and frequencies as well) require a simple division of two pieces of information displayed (in the "probability tree," the non-inverted conditional probabilities can even be taken directly from the lower branches). In contrast, as explicated above, the inversion of conditional probabilities (and thus Bayesian reasoning) requires more complex cognitive operations. Note that formulas (1) and (2a), based on the probability tree or the "probability 2 × 2 table," and formula (2b), based on both frequentistic visualizations, *all* entail more operations than the simple division of two pieces of information.

(6) *2 × 2 tables and tree diagrams in secondary schools*:
Finally, it has to be noted that the 2 × 2 table (with conjoint probabilities), the 2 × 2 table (with frequencies), and the tree diagram (with probabilities) are part of the German secondary school curriculum, whereas the "frequency tree" is not. However, (Bayesian) inferences based on both frequency visualizations seem to be much easier than those based on both probability visualizations (Binder et al., 2015), which brings into question the omnipresent application of the latter in the teaching of statistics. This emphasizes the schools' challenge in teaching the intelligent reading of visualizations (i.e., the facets "read the data," "read between the data," and "read beyond the data" from Curcio, 1989).

Error Strategies Detectable in Tree Diagrams and 2 × 2 Tables

Many statistics educators, but also the psychologists McDowell and Jacobs (2017) in their meta-analysis on Bayesian reasoning, stress the importance of investigating erroneous cognitive algorithms. This, of course, is true for teaching and learning

mathematics in general (e.g., Krauss et al., 2008). But only a few studies have explicitly reported typical incorrect reasoning strategies concerning Bayesian inferences (for some exceptions, see Gigerenzer and Hoffrage, 1995; Steckelberg et al., 2004; Zhu and Gigerenzer, 2006; Eichler and Böcherer-Linder, 2018; Weber et al., 2018).

In order to gain insight into the cognitive problems that people encounter concerning Bayesian inferences and statistical thinking in general, a better understanding of typical errors is required. The few existing classifications of incorrect Bayesian strategies are summarized in **Table 1**. While Gigerenzer and Hoffrage (1995) describe the typical erroneous strategies based on probabilities, Zhu and Gigerenzer (2006) and Eichler and Böcherer-Linder (2018) choose an explanatory approach based on frequencies. To relate all types of errors to our four visualizations (**Figure 1**), we first display both kinds of classifications next to each other (**Table 1**). In doing so, we present the errors based on the notation shown in **Figure 3** (uppercase letters stand for absolute frequencies while lowercase letters represent probabilities). Keep in mind that these letters will later on be used to denote respective areas of interest (AOIs).

Note, however, that the errors reported refer to the typical textual formulations of Bayesian reasoning tasks implemented (see, e.g., the wordings of the mammography problem in the probability and frequency formats in sections Statistical Situations Based on Two Binary Events and Information Formats: Probabilities vs. Frequencies). Gigerenzer and Hoffrage (1995) found the *joint occurrence* to be the most frequent erroneous strategy in Bayesian reasoning. Joint occurrence involves multiplying the *base rate b* and the *sensitivity d* (in frequencies: divide *D* by *A*) without considering the healthy people with positive test results (i.e., *c* and *f*; or correctly dividing *D* by *D*+*F*). According to the same authors, another frequently applied erroneous strategy is the *Fisherian* (or *representative thinking*, according to Zhu and Gigerenzer, 2006) strategy, in which one only takes the sensitivity *d* of the test as the answer (or in terms of frequencies: to calculate *D*/*B*). This error is widespread because it is tempting to confuse P(B|M+) with P(M+|B). Furthermore some participants used another wrong algorithm, which is called *likelihood subtraction* (Gigerenzer and Hoffrage, 1995), meaning erroneously to compute P(M+|B) – P(M+|–B). However, this wrong algorithm predominately occurs in probability versions and is rather unusual for natural frequency versions. A few other participants in that study (Gigerenzer and Hoffrage, 1995) only provided the *base rate b* as the solution of the Bayesian reasoning task, which in frequencies means dividing *B by A* (this error is called *conservatism* by Zhu and Gigerenzer, 2006). The authors also identified the error *evidence-only*, which is the proportion of people with positive test results [i.e., *c* and *f*; or, (*D*+*F*) out of *A*, respectively]. Furthermore, Zhu and Gigerenzer (2006) as well as Steckelberg et al. (2004) reported an error that is documented for frequency versions only, namely *pre-Bayes* (which means to incorrectly divide *B* by *D*+*F*). Finally, some participants also applied the erroneous strategy *correct positive rate/false positive rate* (Steckelberg et al., 2004).

Because visualizations could prevent specific misunderstandings or even block faulty algorithms, it is crucial to reconsider cognitive algorithms with respect to specific visualizations. For instance, the (Fisherian) confusion of P(A|B) with P(B|A) might occur less frequently with a tree diagram (compared to a text-only version) since tree diagrams emphasize the sequential character of the situation more. But even though different visualizations might help for very different reasons, they could also cause new errors that are not listed in **Table 1**. Certain new types of errors might occur according to cognitive load theory (Sweller, 2003) precisely because more information is presented in a tree diagram or in a 2 × 2 table than in a textual version of a Bayesian task. For instance, *E* and *G* or the corresponding probabilities *e* and *g* (cf. **Figure 3**) only appear in visualizations but not in typical wordings, and it is possible for people to erroneously make use of this statistical information in their calculations. It has to be noted that Steckelberg et al. (2004) mention incorrect Bayesian strategies associated with visualizations (tree diagrams and 2 × 2 tables), but do not discuss them in detail. Likewise, possible explanations of the beneficial effect of particular visualizations often remain theoretical (see, e.g., Khan et al., 2015).

For teaching statistics, just as for teaching mathematics in general, it is essential to be an expert on typical errors and on learners' preconceptions (Shulman, 1986, 1987; Krauss et al., 2017). To this end, McDowell et al. (2018) call for a broader methodological approach that can identify typical incorrect Bayesian strategies. Johnson and Tubau (2015) and McDowell and Jacobs (2017) even explicitly suggest eye-tracking analyses of Bayesian reasoning strategies. As educators for future mathematics teachers, we are in addition interested in the pros

TABLE 1 | Correct solution and typical incorrect (Bayesian) strategies according to the correct solution "*D* out of *D* + *F*" in a typical Bayesian reasoning task (according to Gigerenzer and Hoffrage, 1995; Steckelberg et al., 2004; Zhu and Gigerenzer, 2006; Eichler and Böcherer-Linder, 2018).

	Frequencies (with *A*, *B*, *C*, *D*, *E*, *F*, *G)**	**Probabilities (with *b*, *c*, *d*, *e*, *f*, *g*, *h*, *i*, *j*, *k**)**
Correct solution (Bayesian)	***D* out of (*D* + *F*)**	**(*b* · *d*)/(*b* · *d* + *c* · *f*)**
Incorrect algorithm (non-Bayesian)		
Joint occurrence (Gigerenzer and Hoffrage, 1995)	*D* out of *A*	*b* · *d*
Fisherian (Gigerenzer and Hoffrage, 1995)/Representative thinking	*D* out of *B*	*d*
Likelihood subtraction (Gigerenzer and Hoffrage, 1995)	(*D* out of *B*) – (*F* out of *C*)	*d* – *f*
Base rate only (Gigerenzer and Hoffrage, 1995)/conservatism (Zhu and Gigerenzer, 2006)	*B* out of *A*	*b*
Evidence-only (Zhu and Gigerenzer, 2006)	(*D* + *F*) out of *A*	*b* · *d* + *c* · *f*
Pre-Bayes (Steckelberg et al., 2004; Zhu and Gigerenzer, 2006)	*B* out of (*D* + *F*)	Not applicable
Correct positive rate/false positive rate (Steckelberg et al., 2004)	(*D*/*B*) out of (*F*/*C*)	*d*/*f*

* *A, B, C, etc., and b, c, d, etc. represent the pieces of statistical information in the respective visualization (see also* ***Figure 3****).*

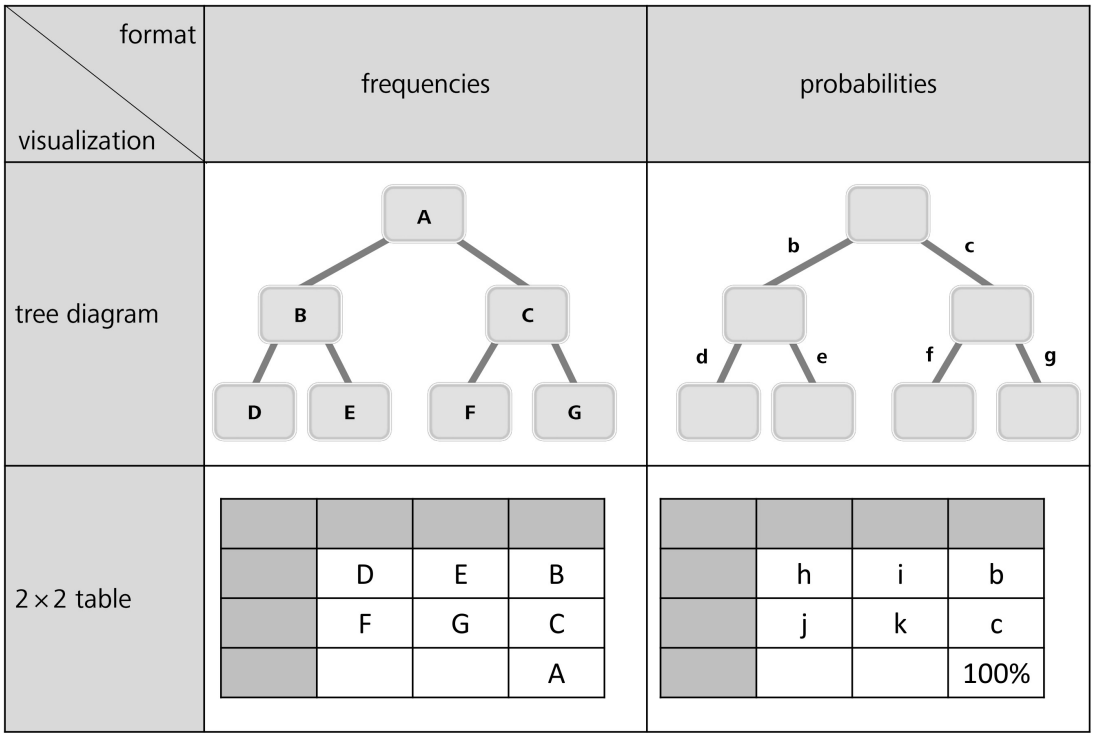


FIGURE 3 | General tree diagrams (above) and 2 × 2 tables (below) provided with frequencies (left) or probabilities (right).

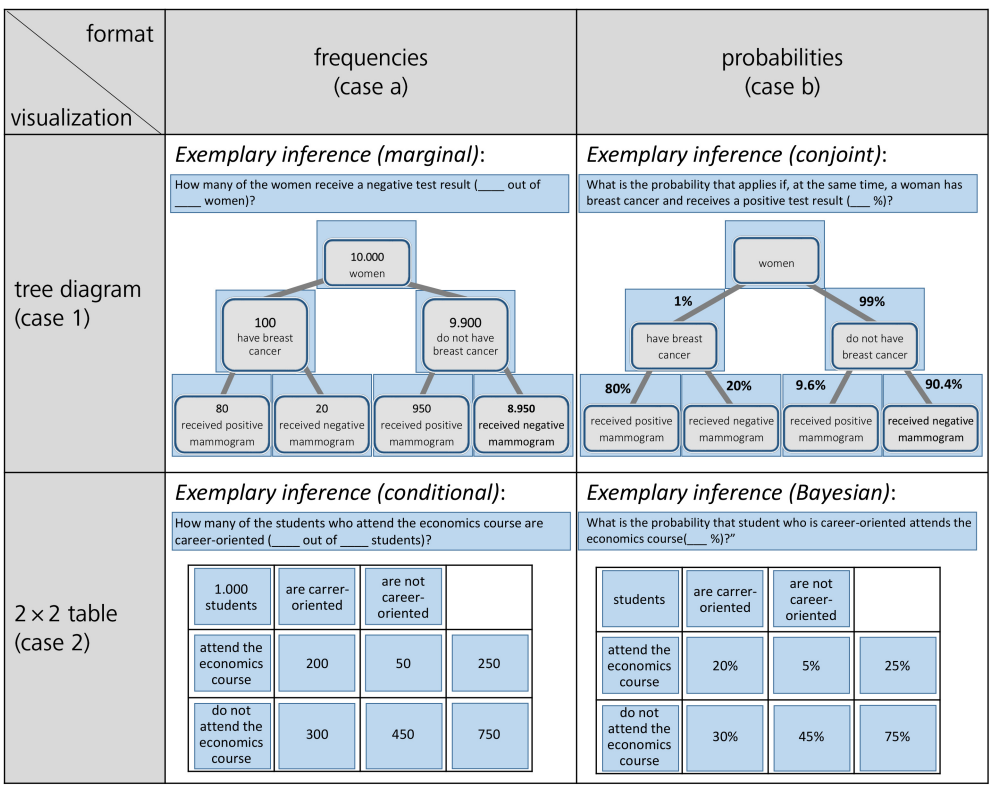


FIGURE 4 | Stimuli for the mammography context and for the economics context (blue-colored AOIs only were included afterwards for the analyses).

and cons of visualizations regarding all 16 possible inferences, especially concerning the most frequently applied visualizations in the (German) context of teaching statistics in secondary schools and universities, namely 2 × 2 tables and tree diagrams.

In the second theoretical section of this paper, we will now focus on the method of eye tracking and how it has been used thus far concerning strategy detection in general but also with respect to statistical reasoning in particular. For this purpose, we introduce the design and results of three studies that are closest to the approach followed in the present article.

EYE TRACKING AS A METHOD FOR ASSESSING STATISTICAL REASONING STRATEGIES

Research Techniques for Identifying Cognitive Processes

Most empirical studies on Bayesian reasoning (or statistical thinking in general) primarily focus on participants' performance rates. However, neither performance rate nor reaction time can fully explain underlying reasoning processes. Verbal reports (or qualitative interviews) might be a path toward an identification of strategies (Robinson, 2001; Smith-Chant and LeFevre, 2003), but participants may have insufficient explicit knowledge to be able to theoretically reflect solution strategies (especially *post hoc*). Therefore, the think-aloud and write-aloud methods (van Someren et al., 1994; for write-aloud protocols on Bayesian reasoning, see Gigerenzer and Hoffrage, 1995) represent an alternative, requiring participants to report on their reasoning strategies simultaneously to their problem solving. However, although this method certainly offers valuable insight into the cognitive strategies that are employed in task processing, it obviously also affects the problem-solving itself.

In contrast, the method of *eye tracking*—a non-invasive measurement of eye movements relative to the head and the visual stimulus—gives a more objective, measurable insight into cognitive and attentional processes involved in, for instance, strategy use or problem solving, without concurrently influencing the process (e.g., Green et al., 2007; Merkley and Ansari, 2010; Huber et al., 2014a). Recording eye movements may therefore be a potential source for capturing thought processes during reasoning and strategy activity. More specifically, and especially with respect to visualizations, it might provide insight into which pieces of information were generally taken into account by a participant and which were not. Thus, eye tracking can be used as a window into cognitive processes that may not be consciously accessible to the participant or apparent to the researcher by task performance (Stephen et al., 2009). Of course, brain-imaging techniques could be a promising additional source of information for combining with techniques like eye tracking within the near future (e.g., see Marian et al., 2003).

Important correlates for cognitive processes during task processing gained by eye tracking are different quantitative and qualitative measures with respect to spatial and temporal features of eye movements that deliver information on eye fixations and saccades. *Fixations* represent the maintaining of the visual gaze on a certain location in the visual field, while fast eye movements from one location to another are called *saccades*. The resulting sequence of fixations and saccades is called a *scan path*, and *dwell time* is the totalized time of all fixations on a given area. In addition, colored *heat maps* aggregate scan paths across different participants, thereby helping researchers to better visualize the relative occurrence of certain scan paths (e.g., see Holmqvist et al., 2011, or **Figures 7–10**).

Eye movements have already been a valuable tool for investigating a number of cognitive domains, including reading (Verschaffel et al., 1992; Meseguer et al., 2002), visual search (Ho et al., 2001), chess (Charness et al., 2001), and problem solving (Epelboim and Suppes, 2001; Knoblich et al., 2001; Thomas and Lleras, 2007). Meanwhile, eye tracking is also being used increasingly within *educational research* (e.g., van Gog and Scheiter, 2010). With respect to *mathematics education*, there are a number of studies that have applied eye movements for innovative findings, for instance regarding arithmetic word problems (e.g., De Corte et al., 1990; Verschaffel et al., 1992; Hegarty et al., 1995), strategies in solving mental addition problems (Verschaffel et al., 1994; Green et al., 2007), fraction comparison (Huber et al., 2014b; Ischebeck et al., 2015; Obersteiner and Tumpek, 2016), number-line estimation strategies (Schneider et al., 2008; Heine et al., 2010; Sullivan et al., 2011), concepts of angles (Schick, 2012), and equation solving (Susac et al., 2014).

Notwithstanding, Verschaffel et al. (2016) point out that "it is remarkable how little researchers in mathematics education have made use of eye tracking so far, particularly for the identification of strategies" (p. 388).

Eye Tracking With Tree Diagrams and 2 × 2 Tables

Only a very few studies have analyzed eye movements during the processing of statistical visualizations like tree diagrams or 2 × 2 tables (especially with respect to Bayesian reasoning tasks), although the method seems well suited to investigating cognitive processes in this domain. In the following, we will describe three relevant eye-tracking studies that deal with at least one of the following aspects: (1) Bayesian reasoning situations, (2) tree diagrams or 2 × 2 tables, and (3) information formats (probabilities and frequencies).

Cohen and Staub (2015) examined wrong strategies in Bayesian reasoning based on purely textual statistical information provided in probabilities. They found that several participants consistently used only one of the three probabilities given in a typical Bayesian reasoning problem (see the respective errors in **Table 1**, e.g., joint occurrence or Fisherian) while other participants used an additive combination of four of the probabilities presented in the tasks (e.g., evidence-only). However, Cohen and Staub (2015) examined only probability versions (but no frequency versions) and did not investigate visualizations in their study.

Lehner and Reiss (2018) analyzed eye movements regarding 2 × 2 tables with absolute numbers (without displaying marginal

distributions). However, they did not ask their participants (students) for probabilities or natural frequencies, but rather for decisions (e.g., "Persons of which sex should be asked if...?"; the absolute numbers of female and male people from two countries were given in the corresponding 2 × 2 table). To answer the implemented questions, students had to focus on one or a combination of two, three or all of the four cells of the visualization. Interestingly, the authors found that the students' gaze durations on single cells differed considerably, with the upper left cell viewed for the most amount of time and the lower right cell for the least amount of time. Moreover, students who were able to solve all of the twelve items with the correct strategy directed their gaze at the lower right cell for a longer period of time than the other participants did. In contrast, students who only solved easier one- or two-cell problems focused for a longer duration on the left column of the table. The authors drew a clear connection between eye movements and (more complex) decision strategies with respect to 2 × 2 tables (Lehner and Reiss, 2018). This research, however, was exclusively focused on 2 × 2 tables containing absolute frequencies and thus tree diagrams or different information formats were not addressed. Furthermore, since no Bayesian reasoning tasks were implemented, the findings cannot be related to **Table 1** of this paper.

Finally, Reani et al. (2017) did indeed investigate the effect of the use of different visualizations with regard to Bayesian reasoning problems. With eye tracking they examined visualizations that were presented in addition to text versions, namely tree diagrams (with frequencies), Euler diagrams (as in **Figure 2**, but with frequencies in the segments of the circles), and icon arrays (without any numerical information). The goal of their study was not primarily to examine whether visualizations facilitate understanding but how students use the presented information. Their eye-tracking data showed that, in line with Lehner and Reiss (2018), participants who answered the presented tasks correctly looked at the stimuli almost twice as long as participants who answered the tasks incorrectly. Regarding frequency trees, they could show that participants looked more intently at information *A* (=total population) than did participants who were presented with a Euler diagram. Conversely, although the performances were identical, regardless of which visualization was used, persons who were shown a Euler diagram viewed information *F* more frequently than persons using a tree diagram (see **Table 1**). However, Reani et al. (2017) analyzed students' eye movements only with respect to frequency-based visualizations. This is relevant to note since in secondary school and university, probability format (instead of frequency format) is usually applied, which is much more at risk for possible errors. Yet only by explicitly investigating 2 × 2 tables and tree diagrams with probabilities can one shed light on the seeming discrepancy between the prominent use and, at the same time, the bad performance attributable to probabilistic visualizations (Binder et al., 2015).

Since (German) students are taught statistics based on 2 × 2 tables and tree diagrams, an eye-tracking analysis systematically comparing both visualizations would seem to be a good source of information that could possibly offer insight regarding underlying cognitive processes (including those that result in errors). As statistics (unfortunately) is usually taught almost exclusively based on probabilities and with probability visualizations, a systematic variation of information format within both visualizations is needed in order to explain the benefit of the format change with respect to these two widely used visualizations.

Present Approach and Research Questions

The present study provides an empirical basis for interpreting eye movements in terms of strategy use concerning statistical situations containing two binary uncertain events. In our approach, we displayed visualizations (tree diagram vs. 2 × 2 table) of such situations. Instead of presenting a complete textual wording, only the requested inferences were shown (above the visualization). On each new screen displaying a certain task in our computer-based experiment, the information format in the visualization changed from probability to frequency (and vice versa), and the requested inference presented above switched between probability and frequency versions accordingly (see **Figure 4** for examples of the final stimuli implemented). In doing so, we examined the strategies of students when they are solving statistical tasks—from easy questions asking for marginal inferences to Bayesian tasks asking for "inverted" conditional inferences (see section Stimuli and Design)—in two different contexts (i.e., mammography context and economics context) by the method of eye tracking, resulting in 20 inferences per participant (see **Table 2** for the design). We investigated how participants looked over those visualizations that comprised the relevant statistical information while answering the questions (within a given time limit).

Our research questions are:

Research question 1:

Which (correct or erroneous) strategies (dependent on visualization type, format, and inference type) used by participants can be detected with the method of eye tracking, and how well can this method predict final performance (i.e., correct or incorrect answer)?

Research question 2:

What can we learn by eye-tracking data about errors made especially in Bayesian reasoning tasks (based on widely applied visualization tools)?

With the first research question (RQ1), we solely want to describe participants' strategies with "classic" quantitative descriptives such as means of solution rates and error types, and compare these results with corresponding heat maps (obtained by scan paths). Thus, in RQ1, we primarily want to check how validly, reliably, and objectively the method of eye tracking can predict the correctness or error type as documented by the purely numerical answer that participants provide as their solution to the task. Since solution strategies and errors are easier to identify with "simple" inferences, we here start with scan paths of non-Bayesian inferences [i.e., marginal, (non-inverted) conditional, and conjoint] regarding RQ1. If scan paths prove to be a valid indicator of participants' reasoning strategies in accordance with RQ1, this method can be used in the second research question

(RQ2) to shed light on (more complicated) Bayesian inferences. Since the effects of visualization and information format have the highest relevance concerning these notoriously difficult problems, in RQ2 we try to explain by eye-tracking data the benefits and problems inherent in both visualizations considering both formats, especially concerning Bayesian inferences.

According to the results of the studies explicated (see section Eye Tracking With Tree Diagrams and 2 × 2 Tables), we expect to find a clear connection between eye movements and certain strategies (see Lehner and Reiss, 2018), which can be found in corresponding spatial and temporal measures. We furthermore expect tree diagrams to be more adequate for some inference types (e.g., conditional probabilities), which might find expression in higher solution rates. Of course, we also expect a replication of the natural frequency effect. With respect to Reani et al. (2017), we expect to find, for instance, that students focus more on areas that are relevant for answering the corresponding questions as compared to other areas (this should apply equally to both information formats), resulting in a higher dwell time and more fixations.

MATERIALS AND METHODS

Participants

A total of 31 adults, all with normal or corrected-to-normal vision, were recruited as a sample for the experiment. Four of these participants had been tested in a pilot study (their eye-tracking data were not included in the present analysis), and the data of three more participants had to be excluded due to their glasses or technical problems. Thus, N = 24 participants (16 female, 8 male) were included in the final analyses. Their mean age was 22.3 (1.6) years, and they ranged from 19 to 26 years of age. The participants were a convenience sample consisting of students from various disciplines at the University of Regensburg (Bavaria, Germany) who were recruited by acquaintance or recommendation. All participants gave their written informed consent and were paid 10 Euro as a representation allowance. While six participants had some unspecific experience with university mathematics due to their studies, the others had only basic mathematical knowledge, and in particular no deeper prior knowledge about (un)conditional probabilities or Bayesian reasoning. Due to their high school education, however, all students were familiar with 2 × 2 tables and tree diagrams containing probabilities, and with 2 × 2 tables containing absolute frequencies, but not with tree diagrams containing frequencies in their nodes (e.g., Binder et al., 2015; Weber et al., 2018).

Eye-Tracking Device

Participants sat in front of a 19-inch computer monitor (with a screen refresh rate of 100 Hz and a resolution of 1280 × 1024 px) at a viewing distance of 70 ± 10 cm. The screen was connected to a remote eye-tracker (iView XRemote RED 250 mobile by SMI) with a sampling rate of 250 Hz. Throughout each trial, the spatial position of each of the observers' eyes ("smart binocular") was sampled running in pupil and corneal reflection mode, resulting in an average spatial accuracy of 0.15°. Participants were asked not to make too many head or body movements, but no device restricted them from moving. Eye movements were calibrated with a five-point, full-screen calibration, both before the experiment began and after a short pause in the middle of the experiment.

Stimuli and Design

Participants were presented two different statistical situations both involving two binary events, namely the *mammography context* and an *economics context* (the latter adapted from Ajzen, 1977; for both contexts, see also Binder et al., 2015). In **Figure 4**, all four combinations of information format and visualization type are displayed (with an exemplary inference; further inferences can be seen in **Table 3**). For each of these two contexts, participants were first asked six non-Bayesian statistical questions—two *marginal*, two (non-inverted) *conditional*, and two *conjoint* inferences, respectively)—in randomized order. After that, they had to answer four (again randomized) *Bayesian* questions in each context, thus resulting in 20 (=2·10) inferences per participant altogether (for the design of the study see **Table 2**; for the implemented infernces see **Table 3**; examples of complete stimuli can be seen in **Figure 4**).

During the administration of each situation (mammography or economics), a large projection of the visualization was shown, with the respective requested inference displayed above the projected image, one after the other. Statistical information on both contexts was given only by this visualization, that is, without additional textual information aside from the question above. To be clear, since both frequency visualizations contain *absolute frequencies*, the term *natural frequencies* strictly speaking refers to the *question format* and not the *information format*. However, absolute frequencies from both visualization types can easily be combined to natural frequencies.

In order to allow familiarization with not only a certain context but also with a specific visualization type, participants always saw a tree diagram for the first ten inferences in the mammography context (factor 1: visualization type). The respective information format within the tree diagram, however, varied randomly, that is, five inferences based on a probability tree and five on a frequency tree (factor 2: information format).

TABLE 2 | Design of the experiment (including 20 resulting inferences per participant).

***N* = 24 students**		**Factor 1: visualization type**	
		Tree diagram (context: mammography problem)	**2 × 2 table (context: economics problem)**
Factor 2: information format	Probabilities	• 1 marginal • 1 conjoint • 1 conditional • 2 Bayesian	• 1 marginal • 1 conjoint • 1 conditional • 2 Bayesian
	Frequencies	• 1 marginal • 1 conjoint • 1 conditional • 2 Bayesian	• 1 marginal • 1 conjoint • 1 conditional • 2 Bayesian

TABLE 3 | Categorization of the four possible inference types (Factor 3) for both contexts.

Factor 3: inference type	Question for	Implemented questions (showing up above the visualizations)
Marginal	$P(B)$	Only in probabilities: "What is the probability that a woman/student receives a positive test result/is career-oriented (___ %)?"
	$P(\neg B)$	Only in natural frequencies: "How many of the women/students receive a negative test result/are not career-oriented (____ out of ____ women/students)?"
Conjoint	$P(A \cap B)$	Only in probabilities: "What is the probability that applies if, at the same time, a woman/a student has breast cancer/attends the economics course and receives a positive test result/is career-oriented (___ %)?"
	$P(A \cap \neg B)$	Only in natural frequencies: "To how many of the women/students does this apply at the same time: They have breast cancer/attend the economics course and receive a negative test result/are not career-oriented (____ out of ____ women/students)?"
	$P(\neg A \cap B)$, $P(\neg A \cap \neg B)$	Not implemented.
Conditional	$P(B\|A)$	Only in natural frequencies: "How many of the women/students who have breast cancer/attend the economics course receive a positive test result/are career-oriented (____ out of ____ women/students)?"
	$P(\neg B\|A)$	Only in probabilities: "What is the probability that a woman/student who has breast cancer/attends the economics course receives a negative test result/is not career-oriented (___ %)?"
	$P(B\|\neg A)$, $P(\neg B\|\neg A)$	Not implemented.
Bayesian	$P(A\|B)$	Only in probabilities: "What is the probability that a woman/student who receives a positive test result/is career-oriented has breast cancer/attends the economics course (___ %)?"
	$P(A\|\neg B)$	Only in natural frequencies: "How many of the women/students who receive a negative test result/are not career-oriented do have breast cancer/attend the economics course (____ out of ____ women/students)?"
	$P(\neg A\|B)$	Only in natural frequencies: "How many of the women/students who receive a positive test result/are career-oriented do not have breast cancer/do not attend the economics course (____ out of ____ women/students)?"
	$P(\neg A\|\neg B)$	Only in probabilities: "What is the probability that a woman/student who receives a negative test result/is not career-oriented doesn't have breast cancer/does not attend the economics course (___ %)?"

Event A: breast cancer or economics course; event B: positive test result or career-oriented.

Afterward, the same procedure was applied for the ten varying inferences (factor 3: inference type) in the economics context, all of which were based on 2 × 2 tables (again, with a randomly varied information format).

In the following, we refer to non-inverted conditional probabilities simply as "conditional probabilities" and to inverted Bayesian conditional probabilities simply as "Bayesian probabilities." The difference between both types of conditional probabilities (and the respective frequencies) as expressed by our visualizations is explained in issue 5 of section Number-Based Visualizations: 2 × 2 Tables and Tree Diagrams.

TABLE 4 | Procedure of the experiment.

Part of experiment	Component (no.)
Introduction	(1) Welcome and introduction.
	(2) Six nature pictures for familiarization with the screen.
Part 1 (visualization: tree diagrams; context: mammography)	(3) Calibration.
	(4) Problem introduction (incl. related narrative) and two example inferences.
	(5) Six non-Bayesian inferences.
	(6) Four Bayesian inferences.
Short pause	(7) /
Part 2 (visualization: 2 × 2 tables; context: economics)	(8) Sequence of components (3)–(6) once again.

*The wordings of each task can be found in **Table 3**.*

- Factor 1: Visualization type: 2 × 2 table (context: mammography problem) vs. tree diagram (context: economics problem)
- Factor 2: Format of statistical information: probabilities vs. absolute frequencies (or natural frequencies in the corresponding question)
- Factor 3: Inference type: marginal vs. conditional vs. conjoint vs. Bayesian (2×).

In **Table 2**, the design is illustrated. Since 24 students participated in the experiment, 480 (=24·20) inferences were made in total, of which 192 (=24·8) were Bayesian inferences. The concrete formulations of the four different types of inferences (displayed above the visualizations) can be found in **Table 3**.

Thus, from all 16 possible questions (see section Statistical Situations Based on Two Binary Events), we posed 10 questions in each context. Therefore, only two out of four conjoint inferences and two out of four non-inverted conditional inferences are missing (see **Table 3**), while the also-missing base rates $P(A)$ and $P(\neg A)$ (unconditional probabilities) were posed as sample questions in the introduction to illustrate the procedure.

Procedure

After a verbal introduction to the experiment that would follow, the procedure began with a short visual introduction [component no. (1), see **Table 4**]; in order to make participants familiar with the device, several nature pictures were shown on the screen (2).

In the first part of the experiment (mammography problem with tree diagrams), initial calibration using cornea reflex was conducted (3). If measurement inaccuracy lay below 0.5° in each direction, the experimental procedure itself began, for which we asked participants to avoid head movements as much as possible. Participants were asked to answer as correctly and as quickly as possible. A time limit of 30 s for each inference was implemented to avoid continuing unspecific, non-target-orientated eye movements.

In both parts of the experiment, the problem contexts were introduced with the help of a short related narrative (e.g., "Imagine you are a reporter from a women's magazine and you want to write an article about breast cancer. You investigate the tests that are conducted in a routine screening

in order to detect breast cancer. The following visualization illustrates the situation."). Then, after participants viewed the situation, they were given two practice trials (4) in order to further familiarize them with the context and both formats (probabilities and frequencies). Both example tasks asked for simple unconditional inferences (i.e., *P(A)* and *P(¬A)* with A being the event "breast cancer" or, in part 2, "economics course"), with one referring to probabilities and the other to frequencies (correct solutions to each were shown afterward). After that, six non-Bayesian inferences followed in random order (5). These six tasks represented a balanced mixture of all possible non-Bayesian tasks (see **Tables 2**, **3**) with respect to format (3 × probabilities, 3 × natural frequencies) and inference type (2 × marginal, 2 × conjoint, and 2 × conditional). If, for instance, one task was given in frequencies [e.g., *P(B|A)*], the other question of the same inference type [$P(\neg B|A)$] was posed in probabilities (see **Table 3**). At the end of part 1, four Bayesian tasks were presented to the participants (6). While two of the four Bayesian questions [$P(A|B)$, $P(\neg A|B)$, $P(A|\neg B)$ or $P(\neg A|\neg B)$] were asked in probabilities, the other two were asked in natural frequencies. Because Bayesian tasks were presented at the end of each part, participants at this stage were already familiar with the context. Thus by this design, purposeless and merely orientating eye movements should have been avoided at least regarding the four final Bayesian inferences in each context. Whenever the format of questions changed the information format in the tree diagram changed correspondingly.

After a short pause (7), the second part of the experiment (8) was conducted parallel to the first part (a calibration was again conducted beforehand). Regarding the inferences concerning the economics context (all ten based on 2 × 2 tables), each participant received the corresponding inference types again systematically varied (see **Tables 2**, **3**).

Participants were assessed individually in a dimly lit room at the University of Regensburg and were asked to speak loudly and communicate their solutions as quickly and as correctly as possible. When they clicked on the F11-key (or when 30 s ran out), the visualization was no longer visible on the screen, but a fixation cross was shown in the middle of the screen; participants then had to immediately state their answer. The experimenter noted down these verbal responses. No feedback was given to the students during the experimental trials. In order to proceed with the next task, participants were required to click the F11-key on the keyboard once again. It was not necessary to use any other key or the computer mouse. In sum, the whole procedure (including introduction, calibrations, pause, etc.) took about 30–40 min.

With respect to traditional coding, a response was classified as a correct answer if either the exact probability or frequency solution was provided or if the indicated probability answer lay within a one percent interval around the correct answer. For instance, in the mammography problem the correct solution to one of the four Bayesian questions is 7.8%, meaning that answers between 7 and 8% were classified as correct (see also Gigerenzer and Hoffrage, 1995).

Data Analysis

While stimuli were presented with the software "Experiment Center 3.0," data analysis of eye movements was conducted using "Suite BeGaze 3.1" (both provided by SMI). To analyze the eye movements, we defined three kinds of "areas of interest" (AOIs) for each screen displaying a task: requested inference (above), concrete information in the visualization, and surrounding white space. **Figure 4** displays four sample (out of 20 different) questions (plus AOIs), one for each visualization × format type. (The AOIs do not belong to the stimuli but were only used for analyses.) Please remind that for each of the four visualizations, five inferences were implemented.

More specifically, the AOIs were fitted around the relevant parts of the screen as follows: With respect to the case of tree diagrams with frequencies (see case 1a in **Figure 4**), both the event and the numerical information were given within the *nodes* of the tree diagram. Here, each of the seven (rectangular) nodes was covered by an equal-sized AOI (each time comprising both number and name of event). In the case of tree diagrams with probabilities (case 1b), numerical information was depicted alongside the *branches* of the diagram; therefore, respective AOIs covered not only the seven nodes (containing the event) but also included the corresponding parts of the branches (containing the respective probability). These AOIs were again equal-sized. In addition, in both cases, the respective inference at the very top of the screen was also covered by an AOI (which was necessarily bigger than the others were). Taken together, eight AOIs covered the whole screen while the rest of the screen was interpreted as a separate area ("whitespace") representing no information. In the case of 2 × 2 tables with either frequencies or probabilities, respectively, the cells themselves were identified as AOIs for both frequencies and probabilities (cases 2a and 2b). Note that regarding 2 × 2 tables in which the name of the event and the corresponding number are not as close to each other as they are in tree diagrams, the four cells containing the events ("attend the economics course," "not attend the economics course," "are career-oriented," and "are not career-oriented") were also covered by an additional AOI. In total, this procedure led to eleven equal-sized AOIs for the 2 × 2 table itself, one additional (bigger) AOI for the requested inference, and the remaining whitespace.

RESULTS

Research Question 1

Regarding the first research question (RQ1)—"Which (correct or erroneous) strategies (dependent on visualization type, format, and inference type) used by participants can be detected with the method of eye tracking, and how well can this method predict final performance (i.e., correct or incorrect answer)?"—we aim at mapping "classic" quantitative statistics on solution and error rates with the corresponding eye-tracking evidence. For doing so, we first discuss solution rates and errors (**Table 5**) that are just based on participants' spoken answers and thus were detectable without eye tracking. Afterward, we report reaction times as well as heat maps regarding participants' scan paths of correct

TABLE 5 | "Classic" descriptives on all inferences.

			Tree diagram			2 × 2 table		
Inference type	**Requested inference**	**Question format**	**Correct answer**	**Performance: pct. correct (#)**	**Incorrect answers (#)**	**Correct answer**	**Performance: pct. correct (#)**	**Incorrect answers (#)**
Marg.	$P(\neg B)$	Nat. freq.	8,970 out of 10,000	95.8% (23 out of 24)	9,000 out of 10,000 (1×)	500 out of 1,000	91.7% (22 out of 24)	50 out of 1,000 (1×), 450 out of 1,000 (1×)
	$P(B)$	Prob.	10.3%	33.3% (8 out of 24)	89.6% (7×), ca. 80% (4×), ca. 1% (3×), 85% (1×), 90% (1×)	50%	83.3% (20 out of 24)	20%, 25%, 35%, 50% (1× each)
Conj.	$P(A \cap \neg B)$	Nat. freq.	20 out of 10,000	50.0% (12 out of 24)	20 out of 100 (7×), 8,950 out of 10,000 (2×), 950 out of 9,900 (2×), 950 out of 10,000 (1×)	50 out of 1,000	79.2% (19 out of 24)	50 out of 250 (3×), 200 out of 1,000 (2×)
	$P(A \cap B)$	Prob.	0.8%	45.8% (11 out of 24)	80% (11×), 1% (1×), 20% (1×)	20%	95.8% (23 out of 24)	80% (1×)
Cond.	$P(B\|A)$	Nat. freq.	80 out of 100	87.5% (21 out of 24)	950/1,030/80 out of 10,000 (1× each)	200 out of 250	75.0% (18 out of 24)	200 out of 1,000 (3×), 200 out of 500 (2×), 300 out of 1,000 (1×)
	$P(\neg B\|A)$	Prob.	20%	83.3% (20 out of 24)	0.2% (2×), 0.02% (1×), 90.4% (1×)	20%	25.0% (6 out of 24)	5% (14×), no answer (2×), 25% (1×), 45% (1×)
Bayes (inverted cond.)	$P(\neg A\|B)$	Nat. freq.	950 out of 1,030	37.5% (9 out of 24)	950 out of 10,000 (=joint occurrence, 5×), 950 out of 9,900 (=Fisherian, 4×), 20 out of 100 (=Fisherian + misread, 2×), no answer (2×), 20 out of 950 (misread, 1×), 8,950 out of 9,030 (misread, 1×)	300 out of 500	79.2% (19 out of 24)	300 out of 1,000 (=joint occurrence) (2×), 200 out of 500 (=misread), 450 out of 500 (=misread), 300 out of 750 (=Fisherian) (1× each)
	$P(A\|B)$	Prob.	≈7.8%	4.2% (1 out of 24)	80% (=Fisherian, 7×), no answer (6×), ca. 90% (="likelihood addition," 3×), 2% (/, 2×), 0.8% (=joint occurrence), 10% and 12% (=evidence only), ca. 20% (=evidence only + miscalculated), 71.4% (=likelihood subtraction) (1× each)	40%	37.5% (9 out of 24)	20% (=joint occurrence) (12×), 66% (=correct positive rate/false positive rate), 75% (=correct positive rate/false positive rate + miscalculated), no answer (1× each)
	$P(A\|\neg B)$	Nat. freq.	20 out of 8,970	41.7% (11 out of 24)	20 out of 10,000 (=joint occurrence, 5×), 20 out of 100 (=Fisherian, 4×), 950 out of 9,900 (=Fisherian + misread, 2×), ca. 100 out of 8,970 (=pre-Bayes, 1×), 80 out of 8,950 (misread, 1×)	50 out of 500	79.2% (19 out of 24)	50 out of 1,000 (=joint occurrence) (2×), 20 out of 500 (=misread), 50 out of 250 (=Fisherian), 50 out of 450 (=correct positive rate/false positive rate) (1× each)
	$P(\neg A\|\neg B)$	Prob.	99.8%	8.3% (2 out of 24)	90.4% (=Fisherian, 8×), (ca.) 90% (=evidence only (2×) or joint occurrence (1×)), ca. 80% (=likelihood subtraction, 3×), (ca.) 95% [=joint occurrence (1×),/(2×)], 98% [=joint occurrence (1×),/(1×)], ca. 97% (=joint occurrence), ca. 96% (/), no answer (1× each)	90%	25.0% (6 out of 24)	45% (=joint occurrence) (13×), 80% (/=miscalculated), 60% (=Fisherian + misread), 30% (=joint occurrence + misread), 22.5% (=/ + miscalculated), about 10% (=correct positive rate/false positive rate) (1× each)

Event A, breast cancer or economics course; event B, positive test result or career-oriented. Likelihood addition means erroneously to add two conjoint probabilities.

answers (**Figures 5**, **6**). Finally, we display the quantitative eye-tracking measures such as dwell time and number of fixations (this time across all participants irrespective of correctness of their answers) for the single AOIs (e.g., *A*, *B*, *C*, etc., and *b*, *c*, *d*, etc.; see **Tables 6**, **7**).

Solution Rates and Errors

Although solution rates are clearly not at the center of the present investigation, they are obviously affected by (correct or incorrect) strategies utilized. **Table 5** presents an overview of solution rates and the absolute frequencies of specific errors for each of the 20 inferences made by the participants. Solution rates vary substantially, ranging from 4.2 to 95.8% across all conditions.

First, in comparing both *visualization types* (factor 1: tree diagram vs. 2 × 2 table), the considerably different solution rates for structurally identical questions—albeit presented with different contexts—immediately catch the eye. Interesting as that is, however, one must keep in mind when comparing quantitative results between both visualization types that the visualization was not randomized in the current study, since the "mammography trees" preceded the "economics 2 × 2 tables" (see Procedure) because the study initially focused on tree diagrams. Thus learning effects might in fact occur. Nonetheless, 2 × 2 tables proved to be more helpful for "marginal" inferences [$P(B)$, $P(\neg B)$], although only for probabilities (tree: 33.3%; 2 × 2: 83.3%) and not for frequencies (tree: 95.8%; 2 × 2: 91.7%). Questions asking for conjunctions [$P(A \cap B)$, $P(A \cap \neg B)$] were also answered at a higher rate of error when accompanied by tree diagrams (freq.: 50.0%; prob.: 45.8%) than they were when accompanied by 2 × 2 tables (freq.: 79.2%; prob.: 95.8%). This is in line with theory since conjunctions only have to be read of the screen in 2 × 2 tables (see section Number-Based Visualizations: 2 × 2 Tables and Tree Diagrams). The opposite applies when it comes to (non-inverted) conditional probabilities [$P(B|A)$, $P(\neg B|\neg A)$], which were answered with a lower rate of error when

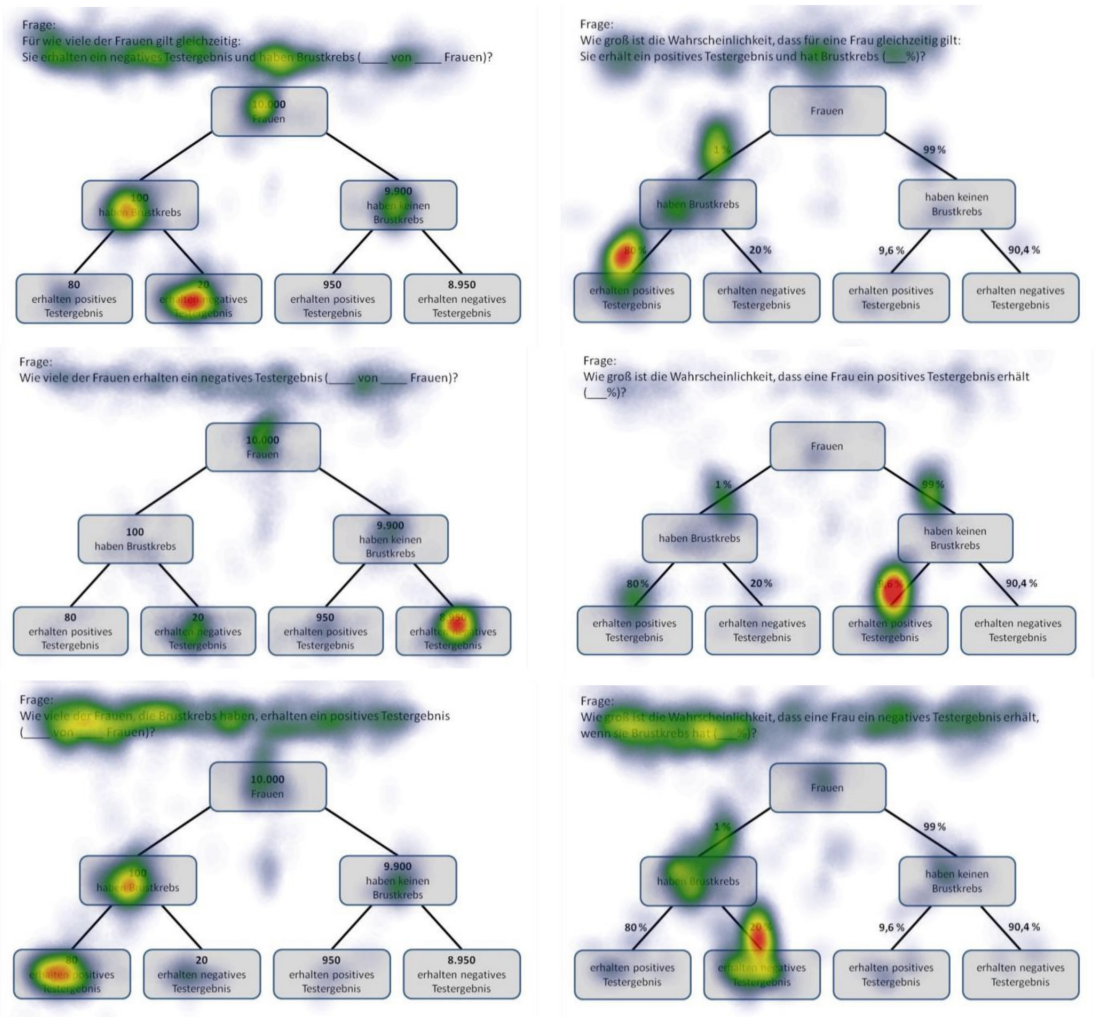


FIGURE 5 | Heat maps of tree diagrams provided with frequencies (left) or with probabilities (right) regarding the following six inferences (from up to below): marginal probabilities, conjoint probabilities, and (non-inverted) conditional probabilities (each only for participants with correct solutions).

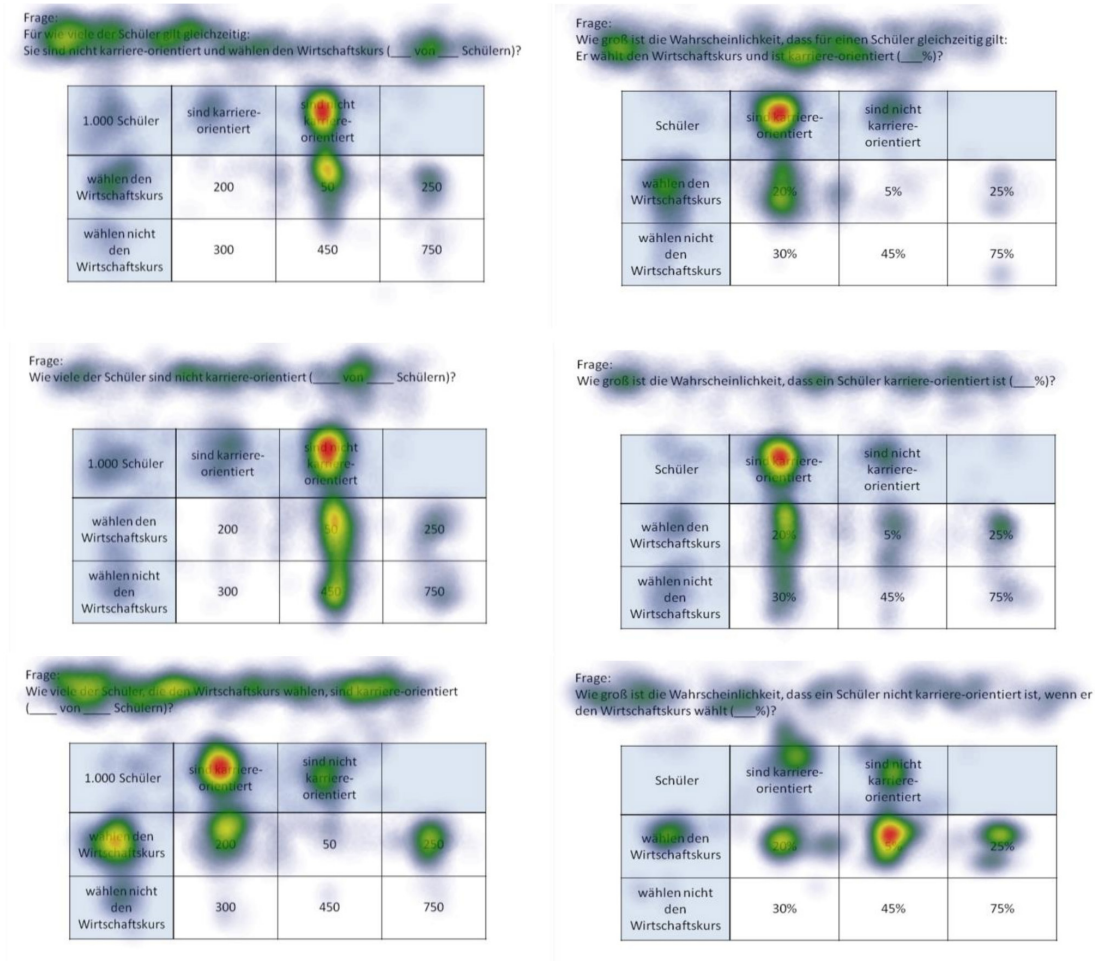


FIGURE 6 | Heat maps of 2 × 2 tables provided with frequencies (left) or with probabilities (right) regarding the following six inferences (from up to below): marginal probabilities, conjoint probabilities, and (non-inverted) conditional probabilities (each only for participants with correct solutions).

accompanied by tree diagrams (freq.: 87.5%; prob.: 83.3%) rather than by 2 × 2 tables (freq.: 75.0%; prob.: 25.0%). Referring to Bayesian inferences (i.e., inverted conditional probabilities), the use of 2 × 2 tables produced either similar or better results than did tree diagrams in relation to all four cases [$P(A|B)$, $P(\neg A|B)$, $P(A|\neg B)$, $P(\neg A|\neg B)$].

Second, regarding *information format* (factor 2: probabilities vs. frequencies), solution rates based on frequencies exceeded those based on probabilities (with one exception) when comparing corresponding questions within tree diagrams (e.g., marginal inferences: freq.: 95.8%; prob.: 33.3%; conjoint inferences: freq.: 50.0%; prob.: 45.8%; conditional inferences: freq.: 87.5%; prob.: 83.3%). The same holds true for the average solution rates of both Bayesian inferences (freq.: 39.6%; prob.: 6.3%). Regarding 2 × 2 tables, similar tendencies were found (marginal inference with freq.: 91.7%; with prob.: 83.3%; conditional inference with freq.: 75.0%; with prob.: 25.0%), except, expectedly, in the case of conjunctions (freq.: 79.2%; prob.: 95.8%). In addition, participants more often solved the two Bayesian tasks correctly in frequency versions than in probability versions (freq.: 79.2 and 79.2%; prob.: 37.5 and 25.0%). When seen in comparison, visualizations presented with frequencies proved to be more easily understandable than those presented with probabilities.

Third, when it comes to different *inference types* (factor 3: marginal vs. conditional vs. conjoint vs. Bayesian), Bayesian tasks, as expected, turned out to be most difficult to solve (39.6% on average across all versions). In probability versions of Bayesian tasks, not only was performance in general relatively low (tree: 6.3%; 2 × 2: 31.3%), but also the kinds of errors that appeared were wide-ranging (see **Table 5**; we will return to the Bayesian inferences in RQ2). In contrast, solution rates of marginal, conjoint, or conditional inferences (across visualization and format: 76.0, 67.7, or 69.8%, respectively) turned out to be substantially higher meaning that these three kinds of inferences are similarly difficult to solve.

Moreover (and pertinent to the focus of the present investigation), **Table 5** exhibits some interesting accumulations of mistakes: Concerning tree diagrams, for instance, some errors regarding non-Bayesian inferences were made by

TABLE 6 | Quantitative performance indicators regarding AOIs in tree diagrams (mammography context).

Requested inference in frequencies	Question	*A* (=10,000)	*B* (=100)	*C* (=9,900)	*D* (=80)	*E* (=20)	*F* (=950)	*G* (=8,950)	White space	Indicators
$P(\neg B)$	1	2	6	5	7	4	8	9	3	Order in sequence
(=8,970 out of 10,000)	5.12 (29.5%)	1.38 (8.2%)	0.34 (1.9%)	0.76 (4.1%)	0.15 (1.0%)	1.69 (10.3%)	0.49 (2.6%)	3.25 (20.5%)	0.47 (3.0%)	Dwell time (in sec./pct.)
(solution rate: 95.8%,	22.2	5.7	1.7	2.6	0.6	6.3	1.6	7.4	3.0	No. of fixations
23 out of 24)	24/24	23/24	14/24	19/24	7/24	22/24	14/24	24/24	23/24	Hit ratio
$P(A \cap \neg B)$	1	3	6	4	9	8	5	7	2	Order in sequence
(=20 out of 10,000)	8.72 (47.0%)	1.06 (6.0%)	1.37 (6.8%)	0.89 (4.9%)	0.21 (1.2%)	1.76 (9.8%)	0.53 (2.9%)	0.36 (1.9%)	0.78 (4.3%)	Dwell time (in sec./pct.)
(solution rate: 50%,	35.9/4.7	4.4/2.7	5.4/3.0	2.6/2.3	0.8/0.4	5.4	2.0	1.1	4.1	No. of fixations
12 out of 24)	24/24	23/24	21/24	13/24	9/24	18/24	10/24	7/24	24/24	Hit ratio
$P(B\|A)$	1	3	5	4	6	7	8	9	2	Order in sequence
(=80 out of 100)	5.86 (46.3%)	0.62 (4.6%)	1.42 (11.0%)	0.31 (2.1%)	1.43 (10.8%)	0.18 (1.3%)	0.26 (2.0%)	<0.01 (0.0%)	0.60 (4.3%)	Dwell time (in sec./pct.)
(solution rate: 87.5%,	25.4	2.9	5.4	0.9	5.1	0.7	1.0	0.0	3.6	No. of fixations
21 out of 24)	24/24	22/24	23/24	10/24	22/24	8/24	4/24	1/24	23/24	Hit ratio
$P(\neg A\|B)$	1	3	5	8	4	7	6	9	2	Order in sequence
(=950 out of 1,030)	7.34 (36.5%)	0.53 (2.8%)	0.75 (3.9%)	1.46 (7.0%)	1.58 (7.4%)	0.54 (2.6%)	3.24 (15.7%)	0.37 (1.8%)	0.70 (3.6%)	Dwell time (in sec./pct.)
(solution rate: 37.5%,	32.3	2.4	2.8	4.2	6.0	2.0	9.9	1.3	4.1	No. of fixations
9 out of 24)	24/24	22/24	18/24	20/24	18/24	16/24	21/24	8/24	22/24	Hit ratio
$P(A\|\neg B)$	1	3	7	4	8	5	6	9	2	Order in sequence
(=20 out of 8,970)	7.52 (34.9%)	0.53 (2.7%)	1.39 (6.5%)	1.09 (4.7%)	0.22 (0.9%)	2.83 (13.2%)	0.87 (3.7%)	2.58 (11.5%)	0.83 (4.1%)	Dwell time (in sec./pct.)
(solution rate: 45.8%,	32.7	2.3	5.3	3.4	0.9	9.5	3.3	7.0	5.0	No. of fixations
11 out of 24)	24/24	21/24	22/24	18/24	7/24	23/24	19/24	20/24	24/24	Hit ratio
Requested inference in probabilities	**Question**	**/(=women)**	***b* (=1%)**	***c* (=99%)**	***d* (=80%)**	***e* (=20%)**	***f* (=9.6%)**	***g* (=90.4%)**	**White space**	**Indicators**
$P(B)$	1	3	4	5	7	8	6	9	2	Order in sequence
(=10.3%)	4.66 (22.1%)	0.55 (2.7%)	1.77 (7.5%)	2.00 (7.9%)	2.48 (10.7%)	0.65 (2.7%)	4.34 (18.5%)	0.32 (1.4%)	1.21 (4.8%)	Dwell time (in sec./pct.)
(solution rate: 33.3%,	20.5	3.1	7.9	7.9	8.4	2.8	13.1	1.4	5.6	No. of fixations
8 out of 24)	24/24	23/24	24/24	20/24	23/24	18/24	21/24	11/24	23/24	Hit ratio
$P(A \cap B)$	1	2	4	7	5	6	9	8	3	Order in sequence
(=0.8%)	7.04 (39.2%)	0.47 (2.4%)	2.72 (13.1%)	0.38 (1.8%)	2.79 (14.5%)	0.64 (3.5%)	0.18 (0.8%)	0.03 (0.1%)	0.88 (4.5%)	Dwell time (in sec./pct.)
(solution rate: 45.8%,	30.0	2.6	10.6	1.3	8.6	2.3	0.7	0.1	4.2	No. of fixations
11 out of 24)	24/24	20/24	22/24	11/24	22/24	16/24	7/24	2/24	24/24	Hit ratio
$P(\neg B\|A)$	1	3	4	6	9	5	7	8	2	Order in sequence
(=20%)	6.07 (43.5%)	0.33 (2.3%)	1.58 (10.7%)	0.56 (3.7%)	0.17 (1.2%)	1.83 (12.8%)	0.25 (1.5%)	0.28 (1.9%)	0.53 (4.0%)	Dwell time (in sec./pct.)
(solution rate: 83.3%,	26.0	1.7	6.5	2.1	0.8	6.3	1.0	0.7	2.9	No. of fixations
20 out of 24)	24/24	19/24	22/24	16/24	9/24	22/24	7/24	5/24	20/24	Hit ratio
$P(A\|B)$	2	3	7	6	4	9	5	8	1	Order in sequence
(≈7.8%)	7.50 (32.1%)	0.45 (2.3%)	2.33 (9.4%)	1.39 (5.2%)	2.99 (12.5%)	0.70 (2.6%)	3.41 (12.1%)	0.13 (0.5%)	0.80 (3.2%)	Dwell time (in sec./pct.)
(solution rate: 4.2%,	32.2	2.0	9.3	5.6	10.2	2.7	10.1	0.6	3.8	No. of fixations
1 out of 24)	24/24	19/24	23/24	19/24	23/24	17/24	17/24	6/24	22/24	Hit ratio
$P(\neg A\|\neg B)$	2	4	8	3	9	7	6	5	1	Order in sequence
(=99.8%)	7.56 (30.9%)	0.32 (1.4%)	0.80 (2.9%)	3.27 (12.8%)	0.19 (0.8%)	2.28 (8.5%)	0.85 (3.5%)	4.16 (17.3%)	0.79 (3.2%)	Dwell time (in sec./pct.)
(solution rate: 37.5%,	32.3	1.7	3.3	10.6	0.8	8.3	3.6	13.4	4.2	No. of fixations
9 out of 24)	24/24	20/24	17/24	24/24	9/24	20/24	21/24	24/24	23/24	Hit ratio

For the first column: event A, breast cancer; event B, positive test result. Gray-colored cells represent AOIs (branches or nodes) relevant to answering the corresponding question correctly. For AOIs denoting A, B, C, etc., and b, c, d, etc., see ***Figure 3****.*

about a third (or more) of all participants [$P(A \cap \neg B)$: "20 out of 100" (7×) instead of "20 out of 1,000"; $P(A \cap B)$: "80%" (11×) instead of "0.8%"; P(B): "89.6%" (7×) instead of "10.3%"]. With Bayesian tasks, participants' wrong answers naturally piled up all the more [e.g., $P(A|B)$: "80%" (=Fisherian) (7×) instead of "0.83%"; $P(\neg A|\neg B)$: "90.4%" (=Fisherian) (8×) instead of "99.8%"]. Second, and very similarly, wrong answers regarding inferences based on 2 × 2 tables indicate common deficient strategies. Most often by far, the (non-Bayesian) conditional probability $P(\neg B|\neg A)$ produced a great number of identical wrong answers [e.g., "5%" (14×) instead of "20%"]. The same holds true for the Bayesian inferences in which two wrong answers in particular (both conforming to joint occurrence and both based on probabilities) appeared to be very tempting [$P(A|B)$: "20%" (12×) instead of "40%"; $P(\neg A|\neg B)$: "45%" (13×) instead of "90%," see **Table 5**]. In all of these cases, analysis of scan paths might reveal a deeper understanding of the specific errors (for details see below).

Reaction Times

Interestingly, the average time it took for participants to reach a solution was not remarkably different for correct or incorrect solutions (in contrast to Reani et al., 2017). In fact, we found differential effects with respect to both visualization types. For instance, regarding the four Bayesian inferences based on tree diagrams, participants who solved the tasks correctly took slightly more time than those who did not [Bayesian inferences with tree diagrams: $M(SD)_{\text{correct}}$ = 23.57(5.78) sec. vs. $M(SD)_{\text{incorrect}}$ = 22.06(7.05) sec.; small effect of d = 0.23 according to Cohen, 1992]. In contrast, with respect to the corresponding four Bayesian inferences based on 2 × 2 tables, the opposite is true: 2 × 2 tables were looked at for a longer period of time by participants who came up with incorrect

TABLE 7 | Quantitative performance indicators regarding AOIs in 2 × 2 tables (economics context).

Requested inference in frequencies	Question	*A* (=1,000)	*B* (=250)	*C* (=750)	*D* (=200)	*E* (=50)	*F* (=300)	*G* (=450)	Event A	Event ¬A	Event B	Event ¬B	White space	Indicators
P (¬*B*)	1	7	12	11	5	2	13	8	9	10	3	4	6	Order in sequence
(=500 out of 1,000)	3.47 (28.4%)	0.34 (3.3%)	0.34 (2.7%)	0.28 (2.2%)	0.13 (1.1%)	1.52 (12.4%)	0.05 (0.3%)	0.92 (7.5%)	0.27 (2.0%)	0.26 (1.9%)	0.54 (4.5%)	1.83 (14.8%)	0.18 (1.4%)	Dwell time (in sec./pct.)
(solution rate: 91.7%,	15.7	1.4	1.4	0.9	0.7	6.5	0.3	3.5	1.1	1.0	1.8	5.7	1.0	No. of fixations
22 out of 24)	24/24	16/24	16/24	12/24	11/24	22/24	6/24	21/24	14/24	10/24	18/24	24/24	15/24	Hit ratio
P (*A* ∩ ¬*B*)	1	7	11	12	4	3	13	9	6	10	2	5	8	Order in sequence
(=50 out of 1,000)	6.79 (44.7%)	0.41 (2.8%)	0.38 (2.5%)	0.11 (0.7%)	0.20 (1.3%)	1.26 (8.5%)	0.02 (0.1%)	0.04 (0.2%)	0.71 (4.4%)	0.19 (1.1%)	0.52 (3.3%)	1.45 (9.3%)	0.26 (1.8%)	Dwell time (in sec./pct.)
(solution rate: 79.2%,	28.0	1.8	1.8	0.5	1.1	5.1	0.1	0.1	2.7	0.6	2.1	4.8	1.4	No. of fixations
19 out of 24)	24/24	16/24	17/24	8/24	14/24	24/24	1/24	4/24	23/24	8/24	16/24	21/24	16/24	Hit ratio
P (*B*|A)	1	4	12	13	5	6	9	10	7	11	3	8	2	Order in sequence
(=200 out of 250)	6.02 (40.4%)	0.43 (3.0%)	0.68 (4.6%)	0.03 (0.1%)	1.23 (8.0%)	0.26 (1.6%)	0.09 (0.6%)	0.01 (0.2%)	1.15 (7.5%)	0.21 (1.1%)	1.50 (9.4%)	0.58 (3.6%)	0.29 (1.9%)	Dwell time (in sec./pct.)
(solution rate: 75%,	26.0	2.0	3.1	0.1	5.6	1.5	0.3	0.1	3.7	0.6	5.7	2.4	1.7	No. of fixations
18 out of 24)	24/24	16/24	19/24	4/24	24/24	18/24	5/24	1/24	22/24	7/24	23/24	17/24	16/24	Hit ratio
P (¬*A*|B)	1	4	11	12	6	2	8	13	10	9	3	7	5	Order in sequence
(=300 out of 500)	6.45 (40.2%)	0.17 (1.2%)	0.13 (0.8%)	0.06 (0.3%)	1.13 (7.1%)	0.32 (1.5%)	0.84 (6.0%)	0.15 (0.6%)	0.59 (3.0%)	0.90 (5.2%)	1.87 (10.7%)	0.52 (2.3%)	0.30 (1.8%)	Dwell time (in sec./pct.)
(solution rate: 79.2%,	27.7	1.0	0.5	0.3	4.7	1.5	3.1	0.6	2.2	2.9	5.9	1.7	1.4	No. of fixations
19 out of 24)	24/24	9/24	8/24	6/24	23/24	13/24	22/24	4/24	16/24	21/24	24/24	13/24	16/24	Hit ratio
P (*A*|¬B)	1	9	13	12	5	3	7	11	8	10	2	4	6	Order in sequence
(=50 out of 500)	7.19 (41.2%)	0.19 (1.0%)	0.18 (0.8%)	0.04 (0.2%)	0.17 (0.9%)	1.82 (10.6%)	0.01 (0.1%)	0.48 (2.8%)	0.95 (5.2%)	0.15 (0.7%)	0.34 (2.0%)	2.62 (13.8%)	0.26 (1.4%)	Dwell time (in sec./pct.)
(solution rate: 79.2%,	30.4	0.7	0.8	0.1	0.8	7.4	0.0	1.9	3.5	0.6	1.6	8.0	1.3	No. of fixations
19 out of 24)	24/24	12/24	8/24	3/24	9/24	22/24	1/24	19/24	20/24	6/24	17/24	22/24	16/24	Hit ratio
Requested inference in probabilities	**Question**	**/ (=100%)**	***b* (=25%)**	***c* (=75%)**	***h* (=20%)**	***i* (=5%)**	***j* (=30%)**	***k* (=45%)**	**Event A**	**Event ¬A**	**Event B**	**Event ¬B**	**White space**	**Indicators**
P (*B*)	1	4	9	13	5	3	7	12	8	11	2	6	10	Order in sequence
(=50%)	4.10 (27.7%)	0.27 (1.2%)	0.78 (3.6%)	0.36 (2.3%)	1.99 (10.9%)	0.78 (4.6%)	0.96 (5.9%)	0.38 (2.1%)	0.57 (2.5%)	0.56 (3.4%)	2.31 (14.5%)	0.74 (3.9%)	0.18 (1.3%)	Dwell time (in sec./pct.)
(solution rate: 83.3%,	18.1	1.3	3.1	1.5	7.9	3.4	3.7	1.3	2.0	1.9	7.7	3.1	1.0	No. of fixations
20 out of 24)	24/24	10/24	16/24	14/24	24/24	21/24	21/24	14/24	16/24	16/24	22/24	18/24	16/24	Hit ratio
P (*A* ∩ *B*)	1	10	13	8	5	2	9	12	6	11	3	7	4	Order in sequence
(=20%)	6.19 (47.5%)	0.08 (0.5%)	0.21 (1.3%)	0.05 (0.4%)	1.28 (9.3%)	0.17 (1.3%)	0.02 (0.2%)	0.01 (0.1%)	0.92 (7.1%)	0.10 (0.7%)	1.56 (11.6%)	0.39 (2.7%)	0.23 (1.8%)	Dwell time (in sec./pct.)
(solution rate: 95.8%,	26.4	0.4	0.7	0.1	4.1	0.9	0.1	0.1	3.3	0.4	5.6	1.4	1.2	No. of fixations
23 out of 24)	24/24	7/24	11/24	2/24	22/24	14/24	2/24	1/24	23/24	6/24	23/24	15/24	15/24	Hit ratio
P (¬*B*|A)	1	5	10	13	3	6	12	11	8	9	2	4	7	Order in sequence
(=20%)	6.71 (40.1%)	0.14 (0.9%)	0.76 (3.2%)	0.03 (0.2%)	0.77 (3.6%)	2.16 (11.1%)	0.06 (0.3%)	0.20 (1.1%)	1.15 (6.6%)	0.29 (1.4%)	0.60 (3.3%)	1.89 (10.9%)	0.25 (1.5%)	Dwell time (in sec./pct.)
(solution rate: 25%,	29.0	0.6	2.8	0.1	2.9	7.5	0.2	0.7	3.5	1.0	2.2	6.4	1.0	No. of fixations
6 out of 24)	24/24	9/24	16/24	2/24	16/24	23/24	3/24	6/24	21/24	10/24	19/24	23/24	14/24	Hit ratio
P (*A*|B)	1	7	10	13	5	2	9	12	8	11	3	6	4	Order in sequence
(=40%)	7.08 (37.4%)	0.18 (0.9%)	0.23 (1.0%)	0.02 (0.1%)	3.02 (14.9%)	0.27 (1.3%)	0.72 (3.1%)	0.02 (0.1%)	1.29 (6.4%)	0.41 (1.9%)	2.24 (11.2%)	0.33 (1.6%)	0.36 (1.7%)	Dwell time (in sec./pct.)
(solution rate: 37.5%,	30.0	1.1	1.0	0.1	8.5	1.3	2.4	0.1	4.5	1.4	7.4	1.4	1.4	No. of fixations
9 out of 24)	24/24	13/24	11/24	1/24	23/24	16/24	16/24	2/24	23/24	15/24	23/24	13/24	16/24	Hit ratio
P (¬*A*|¬B)	1	7	11	12	4	2	13	6	10	9	3	5	8	Order in sequence
(=90%)	7.98 (37.8%)	0.09 (0.5%)	0.06 (0.3%)	0.21 (1.1%)	0.29 (1.3%)	1.61 (7.3%)	0.20 (0.8%)	2.40 (11.6%)	0.30 (1.2%)	1.00 (5.0%)	0.46 (2.1%)	2.14 (10.0%)	0.22 (1.0%)	Dwell time (in sec./pct.)
(solution rate: 25%,	32.0	0.5	0.4	1.0	1.6	5.4	0.9	6.3	1.2	3.0	1.7	6.3	1.0	No. of fixations
6 out of 24)	24/24	6/24	5/24	12/24	15/24	21/24	11/24	21/24	12/24	20/24	15/24	23/24	16/24	Hit ratio

For the first column: event A, economics course; event B, career-oriented. Gray-colored cells represent AOIs (cells) relevant to answering the corresponding question correctly. For AOIs denoting A, B, C, etc., and b, c, d, etc., see ***Figure 3****.*

solutions than by those who gave correct answers [Bayesian inferences with 2 × 2 tables: $M(SD)_{\text{correct}}$ = 17.31(5.78) sec. as compared to $M(SD)_{\text{incorrect}}$ = 20.03(7.69) sec., d = −0.40] (also see Binder et al., unpublished).

Cognitive Strategies Heat Maps Displaying Correct Answers

Before we begin our analysis, we should mention a qualitative aspect that we immediately noticed about participants' scan paths: Participants tended to look back to the requested inference after initially having looked forward to the inference, and after that to the visualization. It seems as if they wanted to make sure that they had understood the requested inference correctly (see also **Tables 6**, **7**). This occurred even more frequently when the question was either difficult (i.e., low solution rate) or the person subsequently answered the question wrongly.

Heat maps can present the scan paths of, for instance, participants who solved the tasks correctly. In **Figure 5**, such heat maps regarding all six non-Bayesian inferences based on tree diagrams are presented. Corresponding heat maps regarding Bayesian inferences (based on tree diagrams or 2 × 2 tables) are displayed in **Supplementary Material**. These colored maps can serve as an indicator for the validity, reliability, and objectivity of the method in general: As can be seen in **Figure 5**, nodes and branches that were relevant for solving the task based on a given tree diagram precisely and distinctly correspond to the areas at which participants looked for the longest period of time. The same holds true with respect to 2 × 2 tables (see **Figure 6**). Taken together, heat maps indicating the most-viewed areas of a stimulus provide a first clue that participants' (individual) viewing areas correspond to their (individual) viewing strategies.

Because in eye-tracking studies it is not possible to present all qualitative results in detail, only heat maps regarding correct solutions are presented here (see **Figures 5**, **6** for all implemented non-Bayesian inferences, **Figures 7–10** for four sample Bayesian inferences, and **Supplementary Material** for the other four Bayesian inferences). Since heat maps in general prove to be valid indicators of participants' focused areas, and because errors are much more relevant concerning Bayesian inferences, we will return to "*Bayesian error* heat maps" in section Research Question 2.

Quantitative Eye-Tracking Analyses of AOIs (Across Correct and Wrong Answers)

Quantitative eye-tracking data refer to the single AOIs, as labeled in **Figure 3** (*A*, *B*, *C*, etc., and *b*, *c*, *d*, etc., respectively). The upper halves of **Table 6** (mammography context) and **Table 7** (economics context) report results regarding nodes or cells of frequency visualizations, and the lower halves those regarding the corresponding AOIs in probability visualizations. Each cell in both tables displays what is known as *performance indicators* that are calculated on average for all participants irrespective of the correctness of their answers, and which are (from top to bottom in each cell) (a) the ordinal number of a certain AOI considered in the sequence (scan paths), (b) the overall dwell time on the respective AOI (in seconds and percentage-wise), (c) the total number of fixations on this AOI, and (d) the hit ratio (i.e., by how many participants the AOI was viewed). In both tables, gray-colored cells represent AOIs that were relevant to answering the corresponding questions, while the other cells were not relevant. For instance, to compute $P(\neg B)$ (correct answer: "8,970 out of 10,000"), one has to add the numbers in the AOIs *E* ("20") and *G* ("8,950") and put the sum in relation to *A* ("10,000"). Because of the small sample size, in the following we present no inference measures (i.e., *p*-values) in favor of qualitative interpretations.

The *order in sequence* is a condensed measure representing the order in which participants scanned the visualization. Considering all of these numbers within a scan path, this measure corresponds to what participants' averaged scan paths look like chronologically. Quite irrespective of whether an AOI is of relevance or not to answer the corresponding question, both visualization types were tendentially viewed for the first time from top to bottom and from left to right [e.g., see $P(B|A)$]. To be clear, the requested inference is considered first. In tree diagrams, the underlying sample (size) is usually viewed after that (which is the AOI *A* for frequency or the AOI "women" for probability versions), while in 2 × 2 tables, participants usually next looked at event B and the upper cells (which are *D* and *E* for frequency or *h* and *i* for probability visualizations).

The *dwell time* represents the time added up of a participant's fixating on a certain AOI, and therefore is necessarily highly correlated with *number of fixations* (see next paragraph). It is not surprising that the AOI that attracted the most attention by far was the requested inference at the top of the screen. Participants spent between 20% and 50% of their time looking at this area. In more detail, both the percentage of time and the absolute time spent on this instruction were especially high for Bayesian questions [e.g., $P(\neg A|\neg B)$] and relatively low for (easier) marginal inferences [e.g., $P(\neg B)$]. This finding indicates that participants needed more time (to grasp and understand the requested inference correctly) the more difficult the inferences were. In addition, AOIs that had to be looked at in order to answer the questions (i.e., gray-colored cells) attracted more attention than those that were irrelevant. With only a few exceptions [e.g., $P(\neg A|\neg B)$ with tree diagrams], the dwell time in the relevant AOIs (gray-colored cells) for any inference was always higher than the dwell time in the irrelevant AOIs.

The *number of fixations* is simply a total of single fixations that occurred in an AOI. As can be seen with respect to both visualizations, the number of fixations was nearly always highest for the AOIs that contained information that was necessary to answering the corresponding question (gray-colored cells). For instance, answering the conditional probability $P(B|A)$, participants spent at least three fixations on the two relevant AOIs (cells *B* and *D*) and almost completely ignored all others. With only one exception [namely, the AOI *f* for $P(\neg A|\neg B)$ in the tree diagram with probabilities], the average number of fixations on the relevant AOIs was always higher than the average number on all of the irrelevant AOIs. These results further indicate that participants process the information in the relevant areas more intensively.

The *hit ratio* represents the proportion of (all 24) participants who looked at the respective AOI. While—not surprisingly—all participants in each instance viewed each task's instructions, some

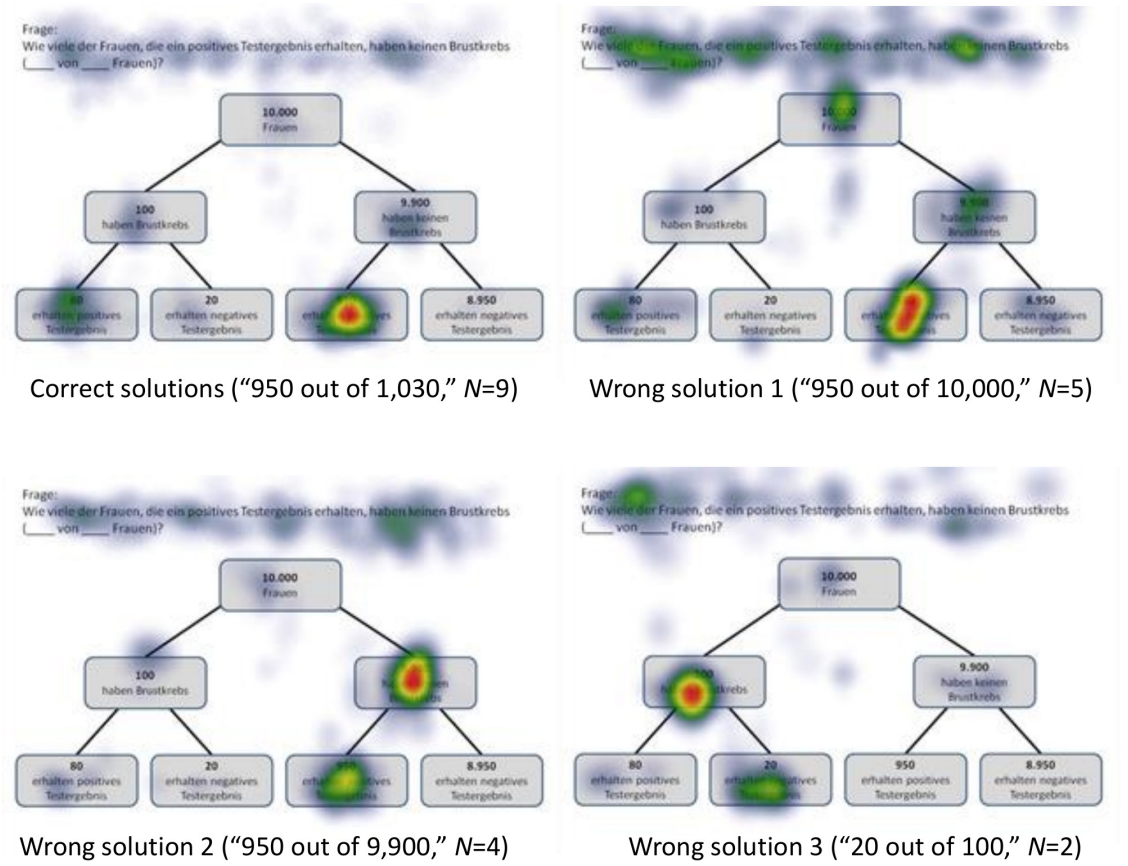


FIGURE 7 | Heat maps regarding the Bayesian inference $P(\neg A|B)$ with a tree diagram (mammography context) with frequencies.

of the irrelevant AOIs were almost completely ignored, which was true especially for the very easy questions [e.g., $P(B|A)$ for tree diagrams or $P(A \cap B)$ for 2×2 tables]. This finding indicates that participants are effectively able to find the relevant information.

In sum, not only heat maps but also performance measures regarding the AOIs (i.e., indicators like order in sequence, dwell time, etc.) obviously provide meaningful evidence of participants' reasoning processes. Both kinds of measures (see **Figures 5**, **6** and **Tables 6**, **7**) can not only be matched with solution and error rates (**Table 5**), but also partly explain erroneous strategies (e.g., Fisherian). This motivates the consideration of these measures with respect to Bayesian inferences in RQ2.

Research Question 2

In the following, we will analyze how solution strategies in Bayesian tasks as evidenced by heat maps and performance indicators (i.e., dwell time, etc.) are impacted by the varying of the two factors *visualization type* and *format*. To do so, we take the two Bayesian inferences $P(\neg A|B)$ and $P(\neg A|\neg B)$ as sample tasks (A reminder: While performance rates of all Bayesian inferences are summarized in the lower half of **Table 5**, performance indicators based on the AOIs of all Bayesian inferences can be found in **Tables 6**, **7**). Heat maps of the two chosen Bayesian inferences, $P(\neg A|B)$ and $P(\neg A|\neg B)$ (both for the correct and the most frequent incorrect strategies), are displayed in **Figures 7**–**10**, whereas the respective heat maps regarding the two unchosen Bayesian inferences, $P(\neg A|B)$ and $P(\neg A|B)$, can be found in **Supplementary Material**. Note that while performance measures of AOIs (**Tables 6**, **7**) again are summarized across all participants' strategies, the heat maps (**Figures 7**–**10** and **Supplementary Material**) explicitly distinguish between correct and incorrect answers.

$P(\neg A|B)$, Based on a Tree Diagram With Frequencies (Mammography Context)

$N = 9$ participants solved the task $P(\neg A|B)$, which asked for a Bayesian inference with frequencies [correct solution: "950 out of 1,030" = "950 out of (950+80)" = "F out of (F+D)"]. As might be expected, participants focused mainly on the relevant AOIs (nodes) D ("80") and F ("950") (but also on A and C; see **Figure 7**). In doing so, they focused much more on F (than on D), which is relevant for both the numerator and the denominator during calculation (besides the mere size of the number). This finding is supported by the high values of number of fixations and dwell time in the corresponding AOIs (although all participants are included, not just those with correct answers).

More interestingly, and of relevance for RQ2, with respect to wrong answer 1 ("950 out of 10,000," $N = 5$), the scan

paths are very similar to those evidenced when selecting the correct answer. For obvious reasons, node *A* ("10,000") was focused on to a greater extent, resulting in a calculation of the "marginal frequency" *P*(*B*) (=error "joint occurrence"). In addition, participants focused more on the question provided above the visualization. With respect to wrong answer 2 ("950 out of 9,900," *N* = 4), participants heavily focused on *C* ("9,900") in addition to *F*, therefore erroneously calculating the conditional probability *P*(*B*|¬*A*) (=Fisherian). Finally, participants giving incorrect answer 3 ("20 out of 100," *N* = 2) focused on the corresponding AOIs *E* ("20") and *B* ("100"), which means that they calculated the "conditional frequency" *P*(*B*|*A*). Obviously, the latter two participants not only executed the wrong calculations, but also misread the question ("receive a negative test result" instead of "receive a positive test result") (=Fisherian).

P(¬*A*|¬*B*), Based on a Tree Diagram With Probabilities (Mammography Context)

The question *P*(¬*A*|¬*B*) required a Bayesian inference with probabilities and was solved correctly by only *N* = 2 participants [correct solution: "99.80%" = 99%·90.4%/(99%·90.4% + 1%·20%) = "(*c*·*g*)/((*c*·*g*) + (*b*·*e*))"]. Participants with the correct answer (all answers between 99 and 100% were classified as correct) focused mainly on the relevant AOIs (branches) *c* ("99%") and *g* ("90.4%") and on the AOIs *b* and *e* (see **Figure 8**), which are relevant for both the numerator and the denominator during calculation. This finding is supported by the maximally high hit ratio (24 out of 24 hits each on AOIs *c* and *g*) and also by the quite high values of dwell time and number of fixations in the corresponding AOIs.

The heat map of all wrong answers (*N* = 15) reveals a particular focus on the AOI *g* ("90.4%"), which was also true for the most prominent wrong answer ["90.4%" (*N* = 8) or "(about) 90%" (*N* = 3)]. Obviously, some of these participants thought that they could simply read on the screen the correct answer from AOI *g* ("90.4%"). Alternatively, some others thought that they had to multiply "90.4%" (AOI *g*) by "99%" (AOI *c*) (≈ 90%). In any case, this is why they more or less ignored the (relevant) AOIs *b* and *e*. While the first incorrect answer represents a conditional probability (=Fisherian), the second corresponds to a conjoint probability [=joint occurrence, or the error "evidence only" = (*c*·*g*)+(*b*·*e*)]. Eye-movement patterns helped to distinguish, for instance, Fisherian from conjoint occurrence errors, even though both mistakes result in nearly the same incorrect answer (e.g., "90.4%" and "ca. 90%," but also "95%" or "98%"). Regarding wrong answer 2 ["(about) 80%," *N* = 3], participants' viewing patterns were quite similar to those of participants who solved the task correctly. Interestingly, as can also be seen in **Figure 8**, their answer, "80%," is obviously not due to AOI *d* ("80%"), which they more or less ignored, nor to the subtraction "90.4–9.6%" (=*g*–*f*). Instead, it seems that they calculated "90.4%–20%" (or "99%–20%") (=likelihood

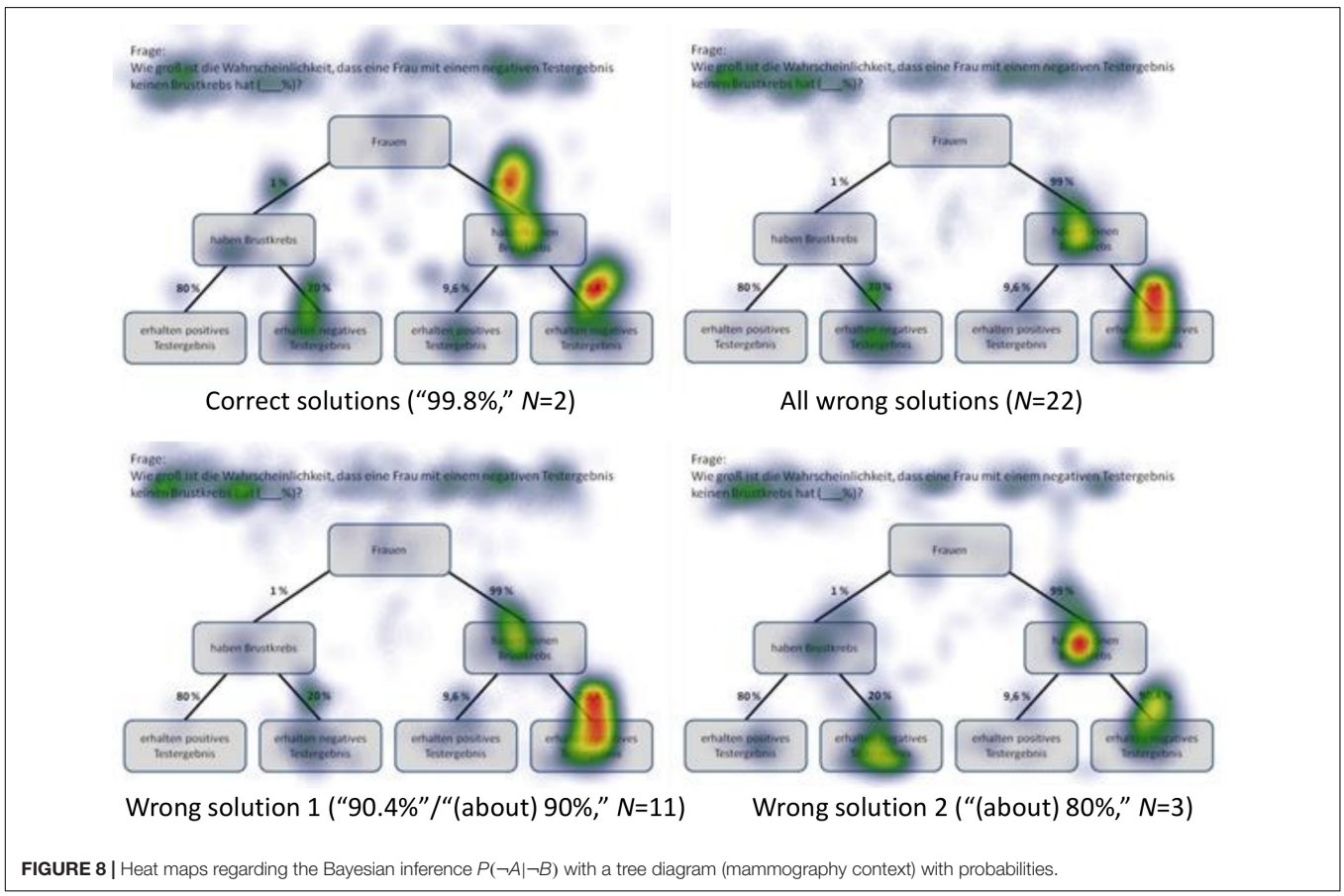


FIGURE 8 | Heat maps regarding the Bayesian inference *P*(¬*A*|¬*B*) with a tree diagram (mammography context) with probabilities.

subtraction). Thus, with respect to RQ2, incorrect reasoning strategies could be detected (only with the help of eye-tracking data) that were not obvious in the given wrong answers itself.

P(¬*A*|*B*), Based on a 2 × 2 Table With Frequencies (Economics Context)

P(¬*A*|*B*) asked for a Bayesian inference with frequencies. It was solved by N = 19 participants [correct solution: "300 out of 500" = 300 out of (300+200) = "D out of (D+F)"]. Participants focused mainly on the relevant AOIs D ("300") and F ("200"), each to a similar extent (see **Figure 9**). In addition, they also focused on the marginal cells "choose the economics course" (event A) and—to an even greater extent—"is career-oriented" (event B), which also finds expression in, for instance, the dwell time and hit ratio on the corresponding AOIs.

With respect to all wrong answers (N = 5), the heat map shows that the marginal cells "choose the economics course" (event A) and (the irrelevant) "not choose the economics course" (event ¬A) were focused on most, both to a very similar extent. However, regarding wrong answer 1 ("300 out of 1,000," N = 2), the corresponding participants' viewing patterns were somehow similar to those of participants with correct solutions, except that the former focused heavily on D ("300"). Also, in contrast to the participants who gave the correct answer, they focused substantially on the marginal cell "1,000 students," which was part of their answer, thus providing a "marginal frequency" (=joint occurrence).

P(¬*A*|¬*B*), Based on a 2 × 2 Table With Probabilities (Economics Context)

Only N = 6 participants solved the question *P*(¬*A*|¬*B*) correctly, which asked for a Bayesian inference based on a 2 × 2 table provided with probabilities [correct solution: "90%" = 45%/(45%+5%) = "k/(k+i)"]. Participants who gave the correct answer focused mainly on the relevant cells k ("45%") and i ("5%") (see **Figure 10**). Interestingly, in doing so, they focused much more on i (than on k), which is relevant only for the calculation of the denominator. This may be because the cell i is positioned between the other two relevant cells. They also focus substantially on the marginal cell "are not career-oriented," which represents the condition ¬B. This finding is supported by the values of number of fixations, dwell time, and hit ratio in the corresponding AOIs (all participants are included).

The heat map of all wrong answers (N = 18) reveals a stronger focus on cell i ("5%") in addition to the corresponding marginal cells ("not choose the economics course" and "are not career-oriented"). The same holds true for the most relevant wrong answer ("45%," N = 13): Obviously, these participants thought that they could read the correct answer from the screen in cell k ("45%"), which is why they more or less ignored the (relevant)

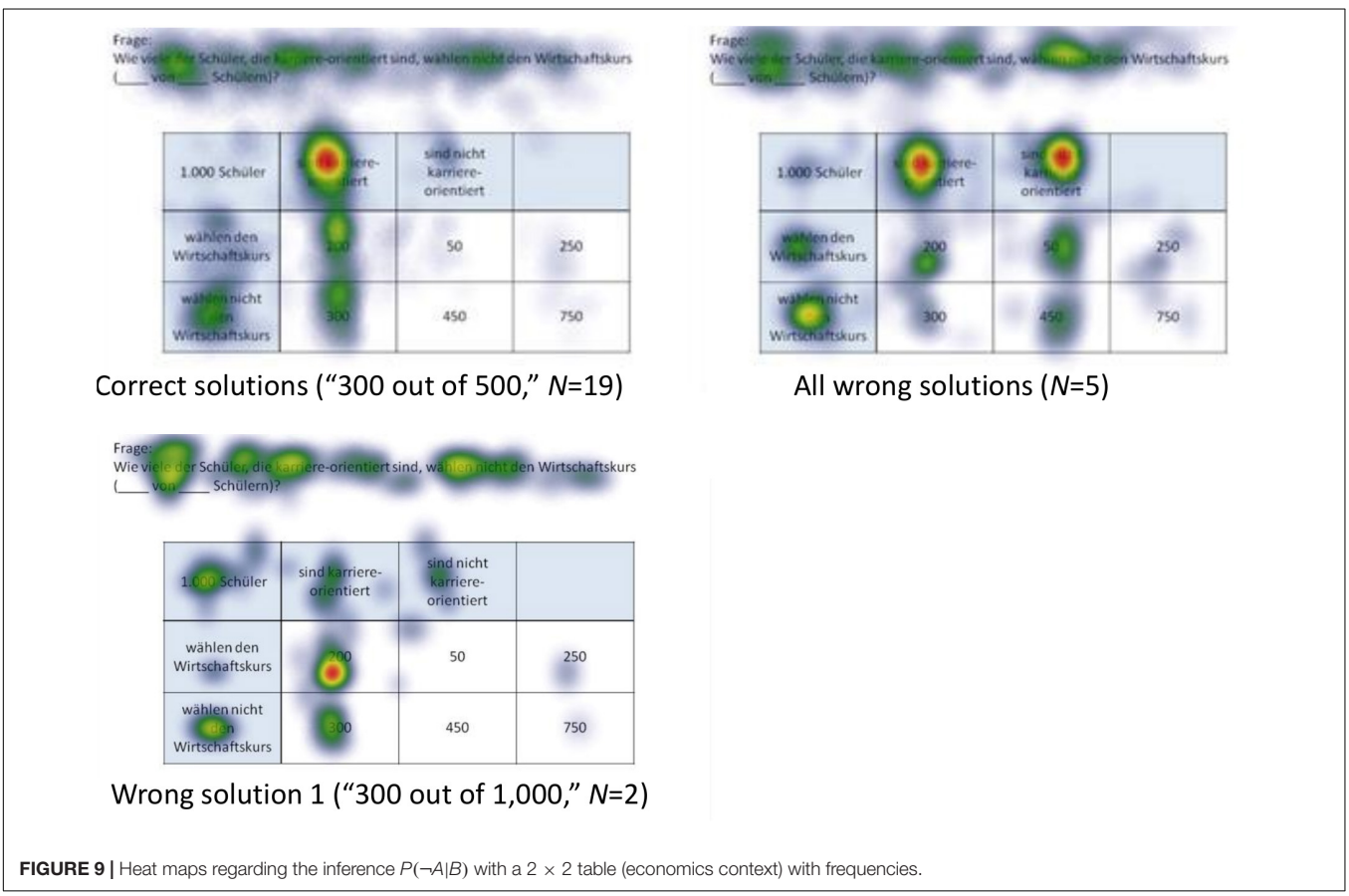


FIGURE 9 | Heat maps regarding the inference *P*(¬*A*|*B*) with a 2 × 2 table (economics context) with frequencies.

FIGURE 10 | Heat maps regarding the inference $P(\neg A|\neg B)$ with a 2 × 2 table (economics context) with probabilities.

cell *i* ("5%"). In doing so, their answer once again erroneously represents a conjoint probability (=joint occurrence).

In sum, the analysis of scan paths by eye tracking revealed, aside from some instances of apparent misreadings, miscalculations, and undefined mistakes, the following recognized errors that can occur in Bayesian tasks (see **Table 8**): The errors "joint occurrence" (in sum: 45×) and "Fisherian" (30×) happened by far the most often. While Fisherian occurred more frequently with tree diagrams (27×) than with 2 × 2 tables (3×), the opposite applies for joint occurrence (tree: 15×; 2×2: 30×). This mismatch is especially due to the high number of joint occurrence errors involving 2 × 2 tables with probabilities (26×), but not involving those with frequencies (4×). All of the other cited errors (e.g., "Pre-Bayes," "likelihood subtraction," etc.; see **Table 1**) could be found in the scan paths and the corresponding answers, but in sum, only quite seldom (15×).

TABLE 8 | Errors per visualization × question/information format for Bayesian inferences.

Visualization	Format: Frequencies	Probabilities
Tree diagram	10× joint occurrence 12× Fisherian 1× Pre-Bayes (in sum: 23× established errors out of 28 errors)	5× joint occurrence 15× Fisherian 3× likelihood subtraction 3× "likelihood addition" 3× evidence only (in sum: 29× out of 45 errors)
2 × 2 table	4× joint occurrence 2× Fisherian (in sum: 6× out of 10 errors)	26× joint occurrence 1× Fisherian 3× correct positive rate/false positive rate (in sum: 30× out of 33 errors)

48 Bayesian inferences per combination.

DISCUSSION

Conclusion

An original feature of this study was the collection of scan paths produced by eye movements during statistical reasoning processes based on tree diagrams and 2 × 2 tables (both provided with probabilities or frequencies). Analyzing students' viewing strategies for solving statistical tasks proved useful as a valid, detailed, and sensitive indicator of participants' reasoning strategies (RQ1). These eye movements provided insight into temporal and spatial distributions of attention during the processing of specific visualizations that are widely applied in the teaching of statistics, not only in Germany but also in many other countries. Since the visualizations provided were presented with either probabilities or frequencies, the participants' solutions also give some hints regarding the benefits and pitfalls (such as provoking particular recognized errors) of different formats in different visualizations. In this way, they call for didactical consequences with respect to the teaching and learning of statistical and especially Bayesian reasoning.

Concerning Bayesian inferences (RQ2), which are intensively examined in cognitive psychology because of their relevance for expert decision-making in various domains, we specifically found the following: Regarding different *visualization types*, tree

diagrams clearly elicit more different kinds of errors than do 2 × 2 tables (see **Table 8**). Viewing patterns (i.e., heat maps) that are essentially a representation of incorrect solutions indicate that 2 × 2 tables especially provoke answers equaling marginal probabilities (or frequencies)—a mistake which is called "joint occurrence" (see **Table 1**). This is logical insofar as 2 × 2 tables, solely due to their structure, display conjoint probabilities in their central cells, thus very much focusing on these probabilities (or frequencies). Moreover, we found only few more established mistakes (i.e., Fisherian, see **Table 8**). Tree diagrams, on the other hand, elicit a variety of incorrect calculations for both formats: We most often encountered "joint occurrence" and "Fisherian," but occasionally "pre-Bayes," "likelihood subtraction," and "evidence-only" as well (see **Table 8**). Thus even though there are obvious benefits of tree diagrams (e.g., see Binder et al., 2015), they more frequently led to different kinds of erroneous calculations in Bayesian questions. One could speculate on whether this is due to their hierarchical structure (contrary to the non-hierarchically structured 2 × 2 tables), which, for example, finds expression in better performances for (non-inverted) conditional inferences for tree diagrams (see **Table 5**). In addition, eye-tracking patterns (i.e., scan paths and heat maps) also revealed that some mistakes were caused by simple misreading (e.g., oversight of a negation) or miscalculations.

Regarding different *formats*, tasks with frequencies were solved to a substantially larger extent than those with probabilities. This result is also reflected in the briefer period of time required to solve frequency tasks (irrespective of whether correct or incorrect answers are compared). Regarding Bayesian inferences, though most participants identified the relevant AOIs for answering a specific inference (as mirrored by dwell time and hit ratio, see **Tables 6**, **7**), neither information format could inhibit the most relevant errors (especially "joint occurrence" and "Fisherian"). The corresponding scan paths and aggregated heat maps (e.g., see **Figures 7**–**10**) support these findings. While participants made only a few different errors in questions posed in natural frequencies, tasks posed in probabilities provoked a greater variety of mistakes, for instance "likelihood addition" (which means erroneously to add two conjoint probabilities) and "evidence only," in addition to some unspecific errors. It seems as if, in contrast to the probability format, the format of frequencies not only reduces errors in general, but also prevents participants from unusual errors (presumably, since the nodes and the cells can very flexibly be combined to multiple insight-fostering natural frequencies).

With respect to different *inference types*, the solution rates of Bayesian tasks expectedly were lower than those of the other inference types. This result also finds expression in the dwell time that participants spent in looking at the instruction: This quantitative measure was especially high for Bayesian questions (and low for marginal inferences). Moreover, we found that participants considered task-relevant AOIs more important than irrelevant AOIs, irrespective of the requested inference type (which is reflected in a higher hit ratio, dwell time, and number of fixations for relevant AOIs). In detail, regarding Bayesian inferences, some typical erroneous Bayesian calculations (see **Table 1**) occurred quite often, while we could detect some others only very rarely (see **Table 8**). Presumably, this finding is due to the given visualizations (rather than mere textual information), which obviously prevents participants from experiencing some (infrequent) misunderstandings.

In sum, and especially with respect to RQ2, the analyses of individual scan paths helped to identify certain strategies, which would not have been possible without eye tracking. For instance, eye tracking helped in interpreting (incorrect) answers that otherwise would have seemed like "nonsense" answers but now could be attributed to misinterpretation, misreading, or miscalculation (see **Table 5,** e.g., for $P(\neg A|\neg B)$ with 2 × 2 tables). Moreover, and especially with respect to probability visualizations in Bayesian tasks, eye-movement analyses revealed that different answers sometimes arise from basically the same errors (see, e.g., $P(\neg A|\neg B)$ with tree diagrams). Conversely, eye tracking helped to distinguish different errors from the same (or very similar) erroneous answers (also see, e.g., $P(\neg A|\neg B)$ with tree diagrams). Furthermore, eye-tracking data revealed that both visualization types are often considered from top to bottom and from left to right (as indicated by the order of sequence), quite similar to the way in which one usually reads a text. Last but not least, participants viewed the requested inferences for quite a long time (and their gaze often returned to them, especially in the case of Bayesian tasks).

The above-mentioned findings, especially the occurrence of very different error distributions with respect to different visualization types and information formats, lead to the following recommendations with respect to the teaching and learning of Bayesian situations: With the results from all inference types (i.e., marginal, conjoint, conditional, and Bayesian) in mind, visualizations should be taught in a more *integrative* and *contrasting* way. This means that, apart from merely showing the visualization (and grasping the relevant information on its own), the "location" of certain information could be explicitly made obvious, for instance by marking the relevant branches or nodes (see Binder et al., 2018). Furthermore, the location of some probabilities or frequencies could explicitly be compared with the location of the same information in other visualizations in order to contrast the different visualizations and information formats (and thus also their advantages and disadvantages). This might lead to a better understanding of which information tree diagrams and 2 × 2 tables display directly (and where), and which inferences cannot be read off but have to be calculated through combining different numbers. In this way, less mixing up of different inference types should occur. Finally, teachers could emphasize the intelligent reading of visualizations (see Curcio, 1989). For instance, if a conditional probability $P(B|A)$ has to be read or computed from a 2 × 2 table, it is somehow more straightforward to focus on the condition first (i.e., on event A, in our study depicted in the columns), and only after that to focus on the corresponding unconditional event (i.e., on event B) in order to compute the correct probability. In tree diagrams, students have to understand that only one "reading direction" is displayed, and thus only one piece of marginal information can be directly read from the tree. In contrast, in double-tree diagrams (e.g., see Wassner, 2004; Khan et al., 2015) both reading directions are displayed at a glance, which is advantageous for

teaching conditional probabilities. In our study, the scan paths of many participants led us to believe that they did not have a deep understanding of how both of the presented visualizations were structured (although they certainly were confronted with them in secondary school).

Limitations of This Study, and Possible Future Research

Qualitative and quantitative eye-movement data and participants' accuracy (i.e., solution rates) provide support for distinguishing among (perhaps unconscious) strategies. Nevertheless, it is necessary to acknowledge that strategies here were derived only indirectly through (aggregated) scan paths (i.e., heat maps), accompanied by the participants' answers. More generally, as it holds true for all eye-tracking studies, it has to be conceded that eye movements and strategy use are by nature related but distinct indicators of thought processes. This is because—similar to gesture—any strategy principally can be performed without the corresponding eye movements as long as the meanings and locations of all the numbers and symbols (e.g., distinct probabilities or frequencies) are known. Future studies, for instance accompanied by retrospective questions to the students intended to help them to figure out their (conscious) strategies, could even more deeply enhance our understanding of participants' thinking.

Moreover, eye-movement data for strategy identification in the domain of mathematical cognition have some general pitfalls (see Verschaffel, 2014): The "process of solving a mathematical problem typically not only consists of an execution phase, but also of an orientation and (possibly) a verification phase" (see Verschaffel et al., 2016, p. 388). Those phases are experimentally hard to separate from each other. In addition, even if one were able to isolate the execution phase, it "frequently may not consist of the straightforward running of a single well-identifiable strategy" (see Verschaffel et al., 2016, p. 388). Taken together, strategies cannot be derived that easily or incautiously. However, we tried to minimize those concerns by keeping the related narrative and the context constant, only changing the corresponding inference (and the information format in the visualization accordingly).

We further acknowledge the limitation that participants were always shown tasks with tree diagrams first, which were then followed by questions with 2 × 2 tables, maybe resulting in a certain learning trajectory from tasks with tree diagrams to those with 2 × 2 tables. Further confounding variables with respect to a comparison of both contexts (and consequently of both visualization types) were somewhat "easier" numbers, the counterintuitive low base rate [i.e., P(A)], and the context itself that might disadvantage tree diagrams as compared to 2 × 2 tables (see also Siegrist and Keller, 2011, for differences in performance of participants in different contexts). For these reasons, comparisons of solution rates and distribution of various mistakes have to be made very cautiously, which might also affect the heterogeneity of wrong answers to some extent. Furthermore, the number of participants was relatively low—although very small case numbers are actually common in eye-tracking studies due to the complexity of their technical implementation. Since quantitative measures obtained can therefore only be interpreted restrictedly, we refrained from inferential statistics. Due to the different structure of both visualization types (hierarchical vs. non-hierarchical) and the location of statistical information (branches or nodes in tree diagrams vs. cells in 2 × 2 tables), both the numbers and the sizes of areas of interest cannot be kept completely comparable, thus in some ways biasing quantitative measures in different conditions. A potential solution to this problem might be to standardize quantitative measures (e.g., fixations) by dividing their number or length by the size and/or number of the respective AOIs.

For future research, it would be interesting to examine the effect of different textual problem formulations on strategies (e.g., for conjoint probabilities, see Hertwig et al., 2008; for conditional probabilities, see partitive vs. non-partitive formulations in Macchi, 2000), since understanding and strategy use are obviously heavily affected by linguistic competencies. In the mammography problem, the more complicated terminology and/or cognitively taxing scenario could also account for the different effects in the different contexts (e.g., Lesage et al., 2013; Sirota et al., 2014a).

Regarding visual aspects, it would also be interesting to analyze the effect of special characteristics of visualizations on viewing patterns. For instance, instead of presenting "normal" tree diagrams or 2 × 2 tables, one could display visualizations with highlighted branches, nodes, or cells in order to figure out the visualizations' effect on participants' eye movements ("signaling principle," see section Number-Based Visualizations: 2 × 2 Tables and Tree Diagrams). Furthermore, it would be interesting to determine whether and how both resources of information (textual and visual) can be integrated or not (and thus shed more light on the "redundancy principle," see section Number-Based Visualizations: 2 × 2 Tables and Tree Diagrams).

Last but not least, the expert-novices paradigm promises some new insights, for example with respect to certain patterns of mistakes: In comparing scan paths and strategies of novices with those of experts, one could perhaps make "learning visible" over time.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of 'Ethikkommission an der Universität Regensburg' with written informed consent from all subjects. Students were informed that their participation was voluntary, and anonymity was guaranteed. After the study participants were debriefed.

AUTHOR CONTRIBUTIONS

GB, KB, and SK contributed by writing the draft of the manuscript. All authors listed have made substantial, direct, and intellectual contribution to the work and approved it for publication. In addition, H-MK and GB recorded the data.

FUNDING

This work was supported by the German Research Foundation (DFG) within the funding program Open Access Publishing, and by a grant from the Academic Research Sabbatical Program of the University of Regensburg, Germany, to GB.

ACKNOWLEDGMENTS

We would like to thank the university students for participating in our study. We are grateful to Maximilian Kölbl for his input in the project by piloting, Barbara Ströhl and Florian Bockes for their assistance with the data collection, and Frances C. Lorié and Patrick Weber for proofreading. In addition, we would like to thank the reviewers for their helpful and constructive comments on previous drafts of this article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00632/full#supplementary-material

REFERENCES

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

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

Binder, K., Krauss, S., Bruckmaier, G., and Marienhagen, J. (2018). Visualizing the Bayesian 2-test case: the effect of tree diagrams on medical decision making. *PLoS One* 13:e0195029. doi: 10.1371/journal.pone.0195029

Böcherer-Linder, K., and Eichler, A. (2017). The impact of visualizing nested sets. An empirical study on tree diagrams and unit squares. *Front. Psychol.* 7:2026. doi: 10.3389/fpsyg.2016.02026

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

Charness, N., Reingold, E. M., Pomplun, M., and Stampe, D. M. (2001). The perceptual aspect of skilled performance in chess: evidence from eye movements. *Mem. Cogn.* 29, 1146–1152.

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

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159.

Curcio, F. R. (1989). *Developing Graph Comprehension*. Reston, VA: N.C.T.M.

De Corte, E., Verschaffel, L., and Pauwels, A. (1990). Influence of the semantic structure of word problems on second graders' eye movements. *J. Educ. Psychol.* 82, 359–365. doi: 10.1037/0022-0663.82.2.359

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

Eichler, A., and Böcherer-Linder, K. (2018). "Categorizing errors in Bayesian situations," in *Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10) Looking Back, Looking Forward*, eds M. A. Sorto, A. White, and L. Guyot (Voorburg: International Statistical Institute).

Eisentraut, F., Ernst, S., Keck, K., Leeb, P., Schätz, U., Steuer, H., et al. (2008). *Delta 10 – Mathematik für Gymnasien [Delta 10 – Mathematics for the Academic School Track]*. Bamberg: CC Buchner.

Epelboim, J., and Suppes, P. (2001). A model of eye movements and visual working memory during problem solving in geometry. *Vis. Res.* 41, 1561–1574.

Fenton, N., Neil, M., and Berger, D. (2016). Bayes and the Law. *Annu. Rev. Stat. Appl.* 3, 51–77.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychol. Rev.* 107, 659–676.

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

Freytag, C., Herz, A., Kammermeyer, F., Kurz, K., Peteranderl, M., Schmähling, R., et al. (2008). *Fokus Mathematik 10 Gymnasium Bayern [Focus on Mathematics 10 for the Bavarian Academic School Track]*. Berlin: Cornelsen Verlag.

Friederichs, H., Ligges, S., and Weissenstein, A. (2013). Using tree diagrams without numerical values in addition to relative numbers improves students' numeracy skills: a randomized study in medical education. *Med. Dec. Mak.* 34, 253–257. doi: 10.1177/0272989X13504499

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.

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

Green, H. J., Lemaire, P., and Dufau, S. (2007). Eye movement correlates of younger and older adults' strategies for complex addition. *Acta Psychol.* 125, 257–278. doi: 10.1016/j.actpsy.2006.08.001

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.

Heine, A., Thaler, V., Tamm, S., Hawelka, S., Schneider, M., Torbeyns, J., et al. (2010). What the eyes already "know": using eye movement measurement to tap into children's implicit numerical magnitude representations. *Infant Child Dev.* 19, 175–186.

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.

Hertwig, R., Benz, B., and Krauss, S. (2008). The conjunction fallacy and the meanings of and. *Cognition* 108, 740–753. doi: 10.1016/j.cognition.2008.06.008

Ho, G., Scialfa, C. T., Caird, J. K., and Graw, T. (2001). Visual search for traffic signs: the effects of clutter, luminance, and aging. *Hum. Factors* 43, 194–207.

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., 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

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

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and Van de Weijer, J. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford: OUP.

Huber, S., Klein, E., Willmes, K., Nuerk, H.-C., and Moeller, K. (2014a). Decimal fraction representations are not distinct from natural number representations – evidence from a combined eye-tracking and computational modeling approach. *Front. Hum. Neurosci.* 8:172. doi: 10.3389/fnhum.2014.00172

Huber, S., Moeller, K., and Nuerk, H. C. (2014b). Adaptive processing of fractions – Evidence from eye-tracking. *Acta Psychol.* 148, 37–48. doi: 10.1016/j.actpsy.2013.12.010

Ischebeck, A., Weilharter, M., and Körner, C. (2015). Eye movements reflect and shape strategies in fraction comparison. *Q. J. Exp. Psychol.* 69, 713–727. doi: 10.1080/17470218.2015.1046464

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

Khan, A., Breslav, S., Glueck, M., and Hornbæk, K. (2015). Benefits of visualization in the mammography problem. *Int. J. Hum. Comput. Stud.* 83, 94–113.

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. doi: 10.1007/978-1-4612-4308-3_27

Knoblich, G., Ohlsson, S., and Raney, E. G. (2001). An eye movement study of insight problem solving. *Mem. Cogn.* 29, 1000–1009. doi: 10.3758/BF03195762

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., et al. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *J. Educ. Psychol.* 100, 716–725. doi: 10.1187/cbe.10-03-0014

Krauss, S., Lindl, A., Schilcher, A., Fricke, M., Göhring, A., Hofmann, B., et al. (eds.). (2017). *FALKO: Fachspezifische Lehrerkompetenzen. Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik [FALKO: Subject Specific Teacher Competences. Conception of Professional Knowledge Test in the Subjects German, English, Latin, Physics, Musical Education, Evangelical Religious Education, and Pedagogy]*. Münster: Waxmann.

Lehner, M. C., and Reiss, K. (2018). Entscheidungsstrategien an vierfeldertafeln: eine analyse mit blickbewegungen [Decision strategies in 2x2 tables: an analysis of eye movements]. *J. Math. Didaktik* 39, 147–170.

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

Macchi, L. (2000). Partitive formulation of information in probabilistic problems: beyond heuristics and frequency format explanations. *Organ. Behav. Hum. Dec. Process.* 82, 217–236.

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

Marian, V., Spivey, M., and Hirsch, J. (2003). Shared and separate systems in bilingual language processing: converging evidence from eyetracking and brain imaging. *Brain Lang.* 86, 70–82.

Mautone, P. D., and Mayer, R. E. (2001). Signaling as a cognitive guide in multimedia learning. *J. Educ. Psychol.* 93, 377–389.

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

Mayer, R. E. (2008). Applying the science of learning: evidence-based principles for the design of multimedia instruction. *Am. Psychol.* 63, 760–769.

McDowell, M., Galesic, M., and Gigerenzer, G. (2018). Natural frequencies do foster public understanding of medical tests: comment on Pighin, Gonzalez, Savadori and Girotto (2016). *Med. Dec. Mak.* 38, 390–399. doi: 10.1177/0272989X18754508

McDowell, M., and Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on bayesian reasoning. *Psychol. Bull.* 143, 1273–1312. doi: 10.1037/bul0000126

Merkley, R., and Ansari, D. (2010). Using eye tracking to study numerical cognition: the case of the ratio effect. *Exp. Brain Res.* 206, 455–460. doi: 10.1007/s00221-010-2419-8

Meseguer, E., Carreiras, M., and Clifton, C. J. R. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Mem. Cogn.* 30, 551–561.

Micallef, L., Dragicevic, P., and Fekete, J. (2012). Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. *IEEE Trans. Vis. Comput. Graph. Inst. Electr. Electron. Eng.* 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

Obersteiner, A., and Tumpek, C. (2016). Measuring fraction comparison strategies with eye-tracking. *ZDM* 48, 255–266.

Operskalski, J. T., and Barbey, A. K. (2016). Risk literacy in medical decision-making. *Science* 352, 413–414.

Pighin, S., Gonzalez, M., Savadori, L., and Girotto, V. (2016). Natural frequencies do not foster public understanding of medical test results. *Med. Dec. Mak.* 36, 686–691. doi: 10.1177/0272989X16640785

Reani, M., Davies, A., Peek, N., and Jay, C. (2017). How do people use information presentation to make decisions in Bayesian reasoning tasks? *Int. J. Hum. Comput. Stud.* 111, 62–77. doi: 10.1037/xlm0000374

Robinson, K. M. (2001). The validity of verbal reports in children's subtraction. *J. Educ. Psychol.* 93, 211–222.

Schick, G. (2012). *Analyse von Eye-Tracking-Daten zur Generierung von Hypothesen über Präkonzepte und Fehlvorstellungen beim Winkelkonzept. [Analysis of Eye-Tracking Data for Generating Hypotheses about Preconceptions and Misconceptions with Respect to Angles]. Beiträge zum Mathematikunterricht 2012*. Münster: WTM.

Schmid, A., Weidig, I., Götz, H., Herbst, M., Kestler, C., Kosuch, H., et al. (2008). *Lambacher Schweizer 10 – Mathematik für Gymnasien Bayern [Lambacher Schweizer 10 – Mathematics for the Bavarian Academic School Track]*. Stuttgart: Ernst Klett.

Schneider, M., Heine, A., Thaler, V., Torbeyns, J., De Smedt, B., Verschaffel, L., et al. (2008). A validation of eye movements as a measure of elementary school children's developing number sense. *Cogn. Dev.* 23, 409–422. doi: 10.1016/j.cogdev.2008.07.002

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

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15, 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harv. Educ. Rev.* 57, 1–22.

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.

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

Sloman, S. A., Over, D., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Dec. Process.* 91, 296–309.

Smith-Chant, B. L., and LeFevre, J.-A. (2003). Doing as they are told and telling it like it is: self-reports in mental arithmetic. *Mem. Cogn.* 31, 516–528.

Spiegelhalter, D., and Gage, J. (2015). What can education learn from real-world communication of risk and uncertainty? *Math. Enthusiast* 12, 4–10. doi: 10.1002/chp.21184

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. *BMC Med. Educ.* 4:13. doi: 10.1186/1472-6920-4-13

Stephen, D. G., Boncoddo, R. A., Magnuson, J. S., and Dixon, J. A. (2009). The dynamics of insight: mathematical discovery as a phase transition. *Mem. Cogn.* 37, 1132–1149. doi: 10.3758/MC.37.8.1132

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)*, eds K. Makar, B. de Sousa, and R. Gould (Voorburg: International Statistical Institute).

Sullivan, J. L., Juhasz, B. J., Slattery, T. J., and Barth, H. C. (2011). Adults' number-line estimation strategies: evidence from eye movements. *Psychon. Bull. Rev.* 18, 557–563. doi: 10.3758/s13423-011-0081-1

Susac, A., Bubiæ, A., Planinic, M., and Palmovic, M. (2014). Eye movements reveal students' strategies in simple equation solving. *Int. J. Sci. Math. Educ.* 12, 555–577.

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

Thomas, L. E., and Lleras, A. (2007). Moving eyes and moving thought: on spatial compatibility between eye movements and cognition. *Psychon. Bull. Rev.* 14, 663–668.

van Gog, T., and Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learn. Instr.* 20, 95–99.

van Someren, M. W., Barnard, Y. F., and Sandberg, J. A. C. (1994). *The Think Aloud Method: a Practical Approach to Modelling Cognitive*. London: Academic Press.

Verschaffel, L. (2014). It's all about strategies, stupid. Invited introduction to the theme "Arithmetic strategies". *Paper Presented at the Expert Meeting on Mathematical Thinking and Learning*, Leiden.

Verschaffel, L., De Corte, E., Gielen, I., and Struyf, E. (1994). "Clever rearrangement strategies in children's mental arithmetic: a confrontation of eye-movement data and verbal protocols," in *Research on Learning and Instruction of Mathematics in Kindergarten and Primary School*, ed. J. E. H. V. Luit (Doetinchem: Graviant Publishing Company), 153–180.

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.

Verschaffel, L., Lehtinen, E., and Van Dooren, W. (2016). Neuroscientific studies of mathematical thinking and learning: a critical look from a mathematics education viewpoint. *ZDM* 48, 385–391.

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

Weber, P., Binder, K., and Krauss, S. (2018). Why can only 24% solve Bayesian reasoning problems in natural frequencies? Frequency phobia in spite of probability blindness. *Front. Psychol.* 9:1833. doi: 10.3389/fpsyg.2018.01833

Woike, J. K., Hoffrage, U., and Martignon, L. (2017). Integrating and testing natural frequencies, naïve Bayes, and fast-and-frugal trees. *Decision* 4, 234–260. doi: 10.1037/dec0000086

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: frequency or 823 nested sets? *Exp. Psychol.* 50, 97–106. doi: 10.1027//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.

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. Dec. 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

PERSPECTIVE
published: 28 August 2018
doi: 10.3389/fpsyg.2018.01550

Bayesian Revision vs. Information Distortion

J. Edward Russo [1,2]*

[1] SC Johnson College of Business, Cornell University, Ithaca, NY, United States, [2] Samuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, NY, United States

The rational status of the Bayesian calculus for revising likelihoods is compromised by the common but still unfamiliar phenomenon of information distortion. This bias is the distortion in the evaluation of a new datum toward favoring the currently preferred option in a decision or judgment. While the Bayesian calculus requires the independent combination of the prior probability and a new datum, information distortion invalidates such independence (because the prior influences the datum). Although widespread, information distortion has not generally been recognized. First, individuals are not aware when they themselves commit this bias. In addition, it is often hidden in more obvious suboptimal phenomena. Finally, the Bayesian calculus is usually explained only with undistortable data like colored balls drawn randomly. Partly because information distortion is unrecognized by the individuals exhibiting it, no way has been devised for eliminating it. Partial reduction is possible in some situations such as presenting all data simultaneously rather than sequentially with revision after each datum. The potential dangers of information distortion are illustrated for three professional revision tasks: forecasting, predicting consumer choices from internet data, and statistical inference from experimental results. The optimality of the Bayesian calculus competes with people's natural desire that their belief systems remain coherent in the face of new data. Information distortion provides this coherence by biasing those data toward greater agreement with the currently preferred position—but at the cost of Bayesian optimality.

Keywords: Bayesian calculus, connectionism, desirability bias, forecasting, information distortion, likelihood updating, rationality, statistical inference

OPEN ACCESS

Edited by:
David R. Mandel,
Defence Research and Development Canada, Canada

Reviewed by:
Norman Fenton,
Queen Mary University of London, United Kingdom
Jean Baratgin,
Université Paris 8, France

***Correspondence:**
J. Edward Russo
jer9@cornell.edu

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

Received: *26 June 2018*
Accepted: *06 August 2018*
Published: *28 August 2018*

Citation:
Russo JE (2018) Bayesian Revision vs. Information Distortion. Front. Psychol. 9:1550. doi: 10.3389/fpsyg.2018.01550

The information needed for nearly all important decisions falls into two categories, values and likelihoods. These decisions typically share two characteristics. First, both values and likelihoods are mainly subjective. Even objective information often requires a subjective evaluation of its decision impact. Second, important decisions usually involve the search for additional information that then drives the revision of the values and likelihoods. While there is no optimizing guidance for revising values, there is for likelihoods. That guidance is the Bayesian calculus[1].

[1]The optimality of the Bayesian calculus presumes a fixed sample space of possible outcomes and their probabilities. In many real situations, this assumption is violated as, over time, additional outcomes become recognized or the probabilities of the original outcomes are altered (Baratgin and Politzer, 2006, 2010). The present claim of Bayesian optimality excludes such changes in the probability space. Respecting the difference between the effect of information to change the probabilities in a stable space and those that change the space itself, we use Bayesian revision for the former and reserve the term updating for the latter.

Bayesian revision combines a prior probability and the diagnostic value of a new datum (i.e., a unit of new information). The calculus of that combination requires that the prior probability and the datum contribute independently to the revised posterior probability. Unfortunately, a common phenomenon of likelihood revision can lead to the violation of that independence assumption.

This phenomenon is the predecisional distortion of information (Russo et al., 1996). It is illustrated by a study of whether to invest in a resort hotel (Russo and Yong, 2011). The sole investment criterion was risk as indexed by the probability of financial failure. Thus, the investment decision depended solely on the likelihood of failure. As information about the hotel was presented, the experimental participants revised the probability of the hotel's financial failure. That is, they repeatedly calculated a revised posterior probability after each new unit of information/datum.

The decision process, or equivalently the revision of posterior probabilities, was tracked by requiring two responses for each datum. The first was the judged diagnosticity of that datum. The second was the posterior probability, that is, the datum-driven revision of the likelihood of investing. The predecisional distortion of information is a bias in the evaluation of the new datum/information toward supporting whichever of the decision options is currently "in the lead." Consider potential investors who are leaning toward investing based on all the data seen so far (and captured by the prior probability). Then information distortion occurs when these investors bias their evaluation of the next datum toward investing. Conversely, if the same investors had been leaning toward not investing, information distortion (ID) is manifest as a biased interpretation of the next datum toward not investing. ID is a specific, process-explicated example of the many phenomena exhibiting a confirmation bias.

The impact of the predecisional distortion of information on the Bayesian calculus is depicted in **Figure 1**. The left panel illustrates the independent contributions of the prior and the datum to the posterior probability. The right panel adds the biasing influence of the prior on the datum. This is the influence of the current/prior leaning toward one option on the evaluation of the next datum/information. In their study of the resort investment decision, Russo and Yong (2011) reported significant information distortion (ID). This finding accords with similar risky decisions studied by DeKay et al. (DeKay et al., 2009, 2011; Glöckner and Herbold, 2011; Miller et al., 2013).

The phenomenon of information distortion (ID) during the revision of probabilities raises several questions. First, is ID widespread enough to affect a substantial number of revision tasks? Second, why has this bias not been recognized (and its consequences for the validity of the Bayesian calculus not been appreciated)? Third, what can be done to eliminate ID? The remainder of this article addresses these three questions. It concludes, first, with a consideration of where ID might undermine applications of the Bayesian calculus and, second, with a comment on the frequent clash between the ideal of normative criteria and the reality of human cognition (e.g., Thaler, 1992) of which the Bayesian calculus vs. ID is only one example.

HOW WIDESPREAD IS INFORMATION DISTORTION?

Reviews by DeKay (2015) and by Russo (2015) report the near universal presence of ID in decisions where the relevant information is acquired over time. Besides studies with college students and MTurk workers, ID has been found in decisions made by auditors, entrepreneurs, physicians, prospective jurors, and sales representatives.

In addition, ID is a systematic function of the prior commitment to the tentatively preferred course of action, as indexed by the prior probability. If that commitment is increased, ID rises in parallel (Polman and Russo, 2012).

ID is also persistent. Occasionally the new information/datum is so anti-leader that the posterior probability reflects a reversal of the leading option. In the above example, this might mean switching the tentative preference from investing to not investing. When such a preference reversal occurs, ID biases the evaluation of new information toward the new leading option, such as toward not investing. (However, see, Carlson et al., 2013, for residual traces of an initial preference).

WHY HAS INFORMATION DISTORTION NOT BEEN RECOGNIZED?

The presence of ID has gone unrecognized for multiple reasons. First, decision makers themselves are unaware that they distort new information. When ID is described to experimental subjects, their estimates of it in their own just-completed decision correlates essentially zero with their actual level of ID (Russo, 2015).

Second, sometimes ID is hidden among other biases. For example, ID is one among many possible causes of the desirability bias in which people overestimate the likelihood of a desired event (Russo and Corbin, 2016).

A third reason for the failure to detect the presence of ID may be peculiar to the canonical Bayesian setting. The familiar demonstrations of the Bayesian calculus have tended to use undistortable data, like contrastingly colored balls drawn randomly from an urn. It is impossible to distort the draw of 3 blue and 7 green balls, however strong may be the prior for one color.

In summary, the difficulty of recognizing ID in likelihood revision seems to have multiple causes. Some causes of ID seem omnipresent, such as the absence of self-awareness. Others operate only in certain situations, such as when conflated with other biases like desirability.

IS THERE REMEDIATION?

Remediation can be achieved, but only partially and in select circumstances. Consider first what may be the most obvious tactic for complete elimination, paying people to be accurate and, therefore, unbiased. (Meloy et al., 2006 see also Engel and Glöckner, 2013) found that incentives increased rather than decreased ID. Further, the negative impact of money held when

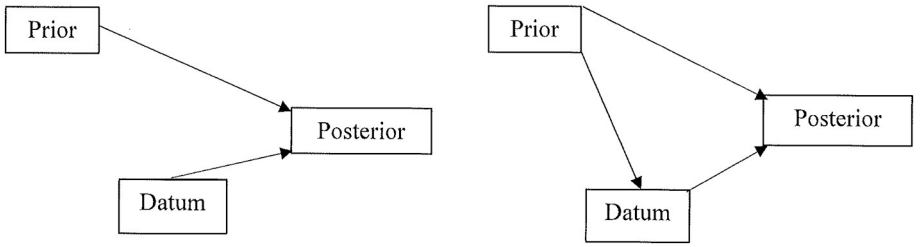


FIGURE 1 | Left panel: valid Bayesian calculus with independent inputs from both the prior probability and the datum to the posterior probability. Right panel: invalid Bayesian calculus because the prior influences the perception and evaluation of the datum, thereby violating their independent contribution to the posterior probability.

people were paid both for the accuracy of the decision and, more importantly, for the accuracy of the evaluation of each datum. (This anomalous result was caused by the positive mood induced by the incentive, Meloy, 2000).

A second potential path to remediation is to identify the cause of ID, which might suggest a method for its elimination. Russo et al. (2008) showed that ID is caused by the general desire for cognitive consistency (Gawronski and Strack, 2012). Specifically, decision makers want the datum (new information) to be consistent with the prior (which reflects all past information). In order to achieve greater consistency they distort the datum in the direction favored by the prior. Unfortunately, knowing that the goal of cognitive consistency drives ID does not reveal a method for ameliorating this bias besides the difficult task of increasing the tolerance for inconsistency. While methods for activating the consistency goal exist (Chaxel and Russo, 2015), those for deactivating it have proved elusive.

In spite of the failure of the above two common paths to amelioration, there have been some partial successes in reducing ID. The first relies on the simultaneous rather than sequential presentation of the data (Carlson et al., 2006). Simultaneous presentation reduces to near zero the delays between data that enable, even encourage, likelihood revision. Of course, such massed presentation does not prohibit spaced revision of the prior. Decision makers can still pause to consider the impact of each new datum before moving to the next information. Nonetheless, the simultaneous presentation of all data seems to inhibit such revision. One reason may be that the time and effort to process all the data at once are likely less than the cumulative total time and effort of several individual revisions.

A second tactic is creating precommitment to the diagnostic value of a datum. That is, each datum is evaluated prior to and, importantly, independent of a particular decision. Carlson and Pearo (2004) showed that if decision makers have knowledge of a datum outside the context of a decision, then ID is reduced almost to zero when that same datum appears during a decision.

Third, decision-making groups exhibit no ID so long as different members maintain opposing positions (Boyle et al., 2012). As long as some members favor one option while other members lean toward another option, there is sufficient debate on the pros and cons of each to suppress ID. This said, most decisions are not made in groups. Further and more worrying, once all members begins to lean toward the same option, ID grows to a level substantially above that of individuals.

In summary, ameliorative tactics are at least partially successful under some circumstances. However, no general strategy for eliminating ID has yet been devised.

THE RISK OF INFORMATION DISTORTION IN APPLICATIONS OF BAYESIAN INFERENCE

An appreciation of the value of Bayesian inference is increasing, as have the number and breadth of its applications. However, with this use of the Bayesian calculus comes the potential risk of contamination by ID. In some environments, such as those with undistortable data, ID can never taint Bayesian inference. However, in other likelihood revision tasks, a recognition of the possible presence of ID may improve the accuracy of Bayesian inference or at least prevent its misapplication.

The increased use of Bayesian inference/methods prompts a consideration of where ID might infect such applications. Three such areas are considered, with no claim to completeness or even representativeness. These are: the forecasting by experts of unique, complex events; the prediction of consumption behaviors from past consumption-related data; and Bayesian approaches to statistical inference. In all three cases, likelihood estimates based on new data/information are essential.

Forecasting by Experts

Although Bayesian methods for likelihood revision have generally not been used where only human judgments can provide a numerical evaluation of a datum, their use is increasingly likely. Consider forecasting, a professional task that has achieved recent success with the identification of "superforecasters" (Mellers et al., 2014, 2015). In the forecasting task, a datum is nearly always a unit of complex information. For instance, if the forecasted event is the reelection of Donald Trump in 2020, a positive datum might be the negotiated end to the Korean conflict of 1950-53. In contrast, a negative datum might be the criminal conviction of one of his inner circle. To apply Bayesian inference, forecasters would have to provide not only an explicit prior probability, as they often do now, but also a numerical judgment of the impact of each new datum, something not routinely required. The Bayesian calculus would then yield the revised posterior probability. In such forecasting, the risk of ID would emerge when experts' commitment to a preferred

event, such as for or against Trump's re-election, biased their evaluation of a new datum (Mandel, 2008).

One familiar forecasting challenge is estimating the likelihood of the success of a new technology[2]. However, what if the experts who must estimate the technology's success are also biased by ID? Consider the example of drug discovery, where pharmaceutical executives must decide whether to pursue the very expensive process of drug development and governmental approval. One of their challenges is that the only credible source of the likelihood of success is the expert scientists who developed the drug. Because they are often committed to its success, they may bias upward their estimated likelihood of eventual success. Yet to whom else can the decision-making executives turn for an informed likelihood of that success?

As typically practiced now, a forecaster need not provide an explicit likelihood ratio for a datum. If that were required, would it reveal a measurable effect of the prior on the evaluation of the datum in the form of ID? That is, might persistent ID in the human experts undermine the superior accuracy of the Bayesian calculus?

Prediction of Consumption

Bayesian models have a substantial history in commerce (e.g., Erdem and Keane, 1996). More specifically, their use by market researchers relies on both past consumption and internet product search to predict future consumption and, more recently, further search (e.g., Ching et al., 2013; Fong, 2017). However, even as these "consumer learning" models become more sophisticated, they tend to be revised between purchases only by the knowledge of what consumer has searched. The evaluation of that acquired information, along with the possible presence of ID, is not included in the models. How much might the predictive accuracy of such models be improved if they accounted for the biasing influence of ID?

Statistical Inference/Analysis

A third and growing domain/area of application of the methods of Bayesian inference is statistical tests of scientific hypotheses. See, for example, the set of papers introduced by Vandekerckhove et al. (2018). As has become well established, using the results from current data to determine when to stop collecting additional data (data-dependent optional stopping) risks invalidating the *p*-values and confidence intervals of classical hypothesis testing. This risk of invalidation has prompted the shift to pre-registering the plan of data collection. Advocates for Bayesian methods claim the elimination of such risks. This claim, in turn, requires the absence of bias in experimenter judgments during the process of inference from collected data.

An analysis of what researchers actually do suggests that this claim may be too strong. For instance, Dunbar (1995, 1999) observed the discussions of research biologists. He found that among the first potential explanations for anomalous data was error in the data collection method (instead of the invalidity of their proffered hypotheses). Surely the same reaction is plausible in the experimental social sciences where data are frequently direct responses from human subjects. Thus, once experimenters who are committed to one hypothesis must judge the validity of their own data, ID may occur.

[2]By the definition of "new," there are no baserates to provide historical evidence of the technology's success. The absence of such baserates precludes the tactic of taking an "outside" view as advocated by Lovallo and Kahneman (2003).

THE RATIONALITY OF BAYESIAN INFERENCE VS. THE MULTIPLE GOALS OF COGNITIVE PROCESSING

The Bayesian calculus belongs to the dominant class of decision theories that rely on the unbiased evaluation of information. Indeed, who would want such a theory if it accommodated rather than rejected a bias like ID? Nonetheless, such theories exist, albeit with descriptive rather than normative status. The most relevant may be connectionist models, which not only accept the ID bias but seem to need it. In these models, new information exerts a bidirectional influence on an existing network of related beliefs. A bidirectional process enables them to accommodate ID as the natural (to these models) influence of a current belief on the evaluation of new information (Holyoak and Simon, 1999; Glöckner and Herbold, 2011). This bidirectional influence contributes to the desired goal of a more coherent and stable system of beliefs as it accommodates to the new information.

Connectionist models do not claim rationality. Nonetheless, the goals of internal coherence and network stability are desirable outcomes of the processing of new information. Thus, an undesirable bias like ID becomes necessary to achieving the desirable ends of coherence and stability (Engel and Glöckner, 2013). Nonetheless, the prominence of connectionist models has tended to obscure the situations where coherence pays the price of tolerating biases like ID.

The descriptive value of connectionist theories coupled with the appeal of the goals that they achieve stands against the normative value of the Bayesian calculus. In tasks where that calculus is needed, such as forecasting, the admittedly desirable goals that drive connectionist dynamics must be sacrificed. Instead, techniques of cognitive engineering need to be developed to counter the natural associative mechanisms that yield ID and other phenomena that compromise the Bayesian calculus. Such help may be found in the techniques that enable superforecasting, such as structured methods for eliciting uncertainty estimates and for statistical reasoning. Two promising examples of structured methods for elicitation are the CHAMPS KNOW training that Mellers et al. (2014) used in the IARPA ACE tournament and Mandel (2015) training of intelligence analysts in Bayesian reasoning using natural sampling trees. Also valuable are disconfirmatory challenges from multiple individuals. When individuals work alone to predict an event of only personal relevance, the techniques of superforecasters may have limited application. Nonetheless, when a task is important enough, such as predicting the success of new technologies like drugs, a team may apply the multiple techniques of superforecasting to achieve Bayesian rationality.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

REFERENCES

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

Boyle, P. J., Hanlon, D., and Russo, J. E. (2012). The value of task conflict to group decisions. *J. Behav. Decis. Mak.* 25, 217–227. doi: 10.1002/bdm.725

Carlson, K. A., Meloy, M. G., and Miller, E. G. (2013). Goal reversion in consumer choice. *J. Consum. Res.* 39, 918–930. doi: 10.1086/666471

Carlson, K. A., Meloy, M. G., and Russo, J. E. (2006). Leaderdriven primacy: using attribute order to affect consumer choice. *J. Consum. Res.* 32, 513–518. doi: 10.1086/500481

Carlson, K. A., and Pearo, L. (2004). Limiting predecisional distortion by prior valuation of attribute components. *Organ. Behav. Hum. Decis. Proc.* 94, 48–59. doi: 10.1016/j.obhdp.2004.02.001

Chaxel, S., and Russo, J. E. (2015). "Cognitive consistency: cognitive and motivational perspectives," in *Neuroeconomics, Judgment, and Decision Making* eds A. W. Evan and F. R. Valerie (New York, NY: Psychology Press), 29–48.

Ching, A. T., Erdem, T., and Keane, M. P. (2013). Learning models: an assessment of progress, challenges, and new developments. *Market. Sci.* 32, 913–938. doi: 10.1287/mksc.2013.0805

DeKay, M. L. (2015). Predecisional information distortion and the self-fulfilling prophecy of early preferences in choice. *Curr. Dir. Psychol. Sci.* 24, 405–411. doi: 10.1177/0963721415587876

DeKay, M. L., Patino-Echeverri, D., and Fischbeck, P. S. (2009). Distortion of probability and outcome information in risky decisions. *Organ. Behav. Hum. Decis. Proc.* 109, 79–92. doi: 10.1016/j.obhdp.2008.12.001

DeKay, M. L., Stone, E. R., and Miller, S. A. (2011). Leader-driven distortion of probability and payoff information affects choices between risky prospects. *J. Behav. Decis. Mak.* 24, 394–411. doi: 10.1002/bdm.699

Dunbar, K. (1995). "How scientists really reason: Scientific reasoning in real-world laboratories," in *The Nature of Insight,* eds R. J. Sternberg and J. Davidson (Cambridge, MA: MIT Press), 365–395.

Dunbar, K. (1999). "How scientists build models: *in vivo* science as a window on the scientific mind," in *Model-based Reasoning in Scientific Discovery*, eds L. Magnani, N. Nersessian, and P. Thagard (Boston, MA: Plenum Press), 89–98.

Engel, C., and Glöckner, A. (2013). Role-induced bias in court: an experimental analysis. *J. Behav. Decis. Mak.* 26, 272–284. doi: 10.1002/bdm.1761

Erdem, T., and Keane, M. P. (1996). Decision-making under uncertainty: capturing dynamic brand choice processes in turbulent consumer goods markets. *Market. Sci.* 15, 1–20. doi: 10.1287/mksc.15.1.1

Fong, N. M. (2017). How targeting affects consumer search. *Manage. Sci.* 63, 2353–2364. doi: 10.1287/mnsc.2016.2447

Gawronski, B., and Strack, F. (2012). *Cognitive Consistency: A Fundamental Principle in Social Cognition*. New York, NY: The Guilford Press.

Glöckner, A., and Herbold, A. K. (2011). An eye-tracking study on information processing in risky decisions: evidence for compensatory strategies based on automatic processing. *J. Behav. Decis. Mak.* 24, 71–98. doi: 10.1002/bdm.684

Holyoak, K. J., and Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *J. Exp. Psychol. Gen.* 128, 3–31. doi: 10.1037/0096-3445.128.1.3

Lovallo, D., and Kahneman, D. (2003). Delusions of success. *Harv. Bus. Rev.* 81, 56–63. doi: 10.1225/R0307D

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

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., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., et al. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspect. Psychol. Sci.* 10, 267–281. doi: 10.1177/1745691615577794

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

Meloy, M. G. (2000). Mood-driven distortion of product information. *J. Consum. Res.* 27, 345–359. doi: 10.1086/317589

Meloy, M. G., Russo, J. E., and Miller, E. G. (2006). Monetary incentives and mood. *J. Market. Res.* 43, 267–275. doi: 10.1509/jmkr.43.2.267

Miller, S. A., DeKay, M. L., Stone, E. R., and Sorenson, C. M. (2013). Assessing the sensitivity of information distortion to four potential influences in studies of risky choice. *Judgm. Decis. Mak.* 8, 662–677. Available online at: https://econpapers.repec.org/RePEc:jdm:journl:v:8:y:2013:i:6:p:662-677

Polman, E., and Russo, J. E. (2012). Reconciling competing beliefs during decision making. *Organ. Behav. Hum. Decis. Proc.* 119, 78–88. doi: 10.1016/j.obhdp.2012.05.004

Russo, J. E. (2015). "The Predecisional Distortion of Information," in *Neuroeconomics, Judgment, and Decision Making,* eds A. W Evan and F. R Valerie (New York, NY: Psychology Press), 91–110.

Russo, J. E., and Corbin, J. (2016). Not by desire alone: the role of cognitive consistency in the desirability bias. *Judg. Dec. Makin.* 11, 449–459. Available online at: https://econpapers.repec.org/RePEc:jdm:journl:v:11:y:2016:i:5:p:449-459

Russo, J. E., Kurt, A. C., Margaret, G. M., and Kevyn Y. (2008). The goal of consistency as a cause of information distortion. *J. Exp. Psychol. Gen.* 137, 456–470. doi: 10.1037/a0012786

Russo, J. E., Medvec, V. H., and Margaret, G. M. (1996). The distortion of information during decisions. *Organ. Behav. Hum. Decis. Proc.* 66, 102–110. doi: 10.1006/obhd.1996.0041

Russo, J. E., and Yong, K. (2011). The distortion of information to support an emerging assessment of risk. *J. Economet.* 162, 132–139. doi: 10.1016/j.jeconom.2010.07.004

Thaler, R. L. (1992). *The Winner's Curse: Paradoxes and Anomalies of Economic Life*. New York, NY: Free Press.

Vandekerckhove, J., Rouder, J. N., and Kruschke, J. K. (2018). Editorial: bayesian methods for advancing psychological science. *Psych. Bull. Rev.* 25, 1–4. doi: 10.3758/s13423-018-1443-8

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 June 2018
doi: 10.3389/fpsyg.2018.00903

Metacognitive Myopia in Hidden-Profile Tasks: The Failure to Control for Repetition Biases

Klaus Fiedler, Joscha Hofferbert and Franz Wöllert*

Department of Psychology, Heidelberg University, Heidelberg, Germany

The failure to exploit collective wisdom is evident in the conspicuous difficulty to solve hidden-profile tasks. While previous accounts focus on group-dynamics and motivational biases, the present research applies a metacognitive perspective to an ordinary learning approach. Assuming that evaluative learning is sensitive to the frequency with which targets are paired with positive versus negative attributes, selective repetition of targets' assets and deficits will inevitably bias the resulting evaluations. As selective repetition effects are ubiquitous, metacognitive monitoring and control functions are required to correct for repetition biases. However, three experiments show that metacognitive myopia prevents judges from correction, even when explicitly warned to ignore selective repetition (Experiment 1), when same-speaker repetitions rule out social validation (Experiment 2) and when blatant debriefing enforces superficial corrections (Experiment 3). For a comprehensive understanding of collective judgments and decisions, it is essential to take metacognitive monitoring and control into account.

Keywords: hidden profiles, repetition bias, group decision making, meta-cognitive myopia, monitoring and control

OPEN ACCESS

Edited by:
David R. Mandel,
Defence Research and Development Canada, Canada

Reviewed by:
Andreas Mojzisch,
University of Hildesheim, Germany
Ans Vercammen,
Imperial College London, United Kingdom

****Correspondence:***
Klaus Fiedler
klaus.fiedler@psychologie.uni-heidelberg.de

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

Received: *18 February 2018*
Accepted: *17 May 2018*
Published: *05 June 2018*

Citation:
Fiedler K, Hofferbert J and Wöllert F (2018) Metacognitive Myopia in Hidden-Profile Tasks: The Failure to Control for Repetition Biases. Front. Psychol. 9:903. doi: 10.3389/fpsyg.2018.00903

INTRODUCTION

Democratic societies rely on the belief that arduous tasks that exceed individual persons' capacity can be managed collectively. Performance and motivation can be enhanced if the overall workload is divided. However, for many judgment and decision problems – such as health risk assessment or personnel selection – the need to coordinate and integrate collective efforts creates a serious difficulty. Information can vary in trustworthiness and validity, arguments may be redundant or in conflict, and individual opinions may rely on different sources and sample sizes. Still, in democratic societies, virtually all important decisions are made collectively.

Despite the trust in the superiority of collective knowledge and in the wisdom of crowds (Surowiecki, 2004; Mannes et al., 2014), several decades of empirical research have drawn a rather pessimistic picture. Collective brainstorming was shown to decrease productivity (Diehl and Stroebe, 1987), group discussion can cause polarization and over-statement (Brauer et al., 1995; McCauley, 1998), and others' advice is not utilized appropriately (Yaniv et al., 2009).

Conspicuous Evidence From Hidden-Profile Tasks

Research on hidden profile-tasks illuminates this failure to exploit the potential advantage of collective wisdom (Stasser and Titus, 1985, 2003; Lu et al., 2012; Schulz-Hardt and Mojzisch, 2012). In this paradigm, part of the information about decision options (applicants, products) is shared by everybody, while other, unshared information is exclusively available to single individuals (see **Table 1**). Although Candidate A (six positive and three

negative attributes) is clearly superior to Candidate B (three positive and six negative attributes), the subset of information available to all three individual judges J1, J2, and J3 favors B (three positive; two negative) over A (two positive; three negative). This is possible because A's few deficits and B's few assets are shared (dark gray) whereas A's many assets and B's many deficits are unshared (light gray). If all three judges follow their individually learned preferences, they will agree on a wrong decision. The only chance to uncover the hidden profile seems to be the collective exchange all raw arguments about all candidates' assets and deficits. However, a growing body of evidence shows that people rarely manage to transcend their individual perspective and to identify a hidden profile (Lu et al., 2012; Schulz-Hardt and Mojzisch, 2012). Several explanations that have been offered for this persistent deficit converge on emphasizing group-dynamic influences and social motives.

Most prominent accounts focus on a shared-information bias. Shared arguments are more likely to be mentioned and repeated in group discussions than unshared arguments (Stasser et al., 1989; Larson et al., 1994; Mesmer-Magnus and DeChurch, 2009), for two reasons. First, shared arguments are known by more than one discussant and are therefore more likely to be mentioned by at least one discussant than unshared arguments (Larson et al., 1994; Larson and Harmon, 2007). Second, shared arguments are socially rewarding and serve to enhance one's self-esteem (Wittenbaum et al., 1999). Complementing the shared information bias is a bias to discuss (Dennis, 1996; Faulmüller et al., 2012) or to believe in the validity of preference-consistent arguments (Edwards and Smith, 1996; Greitemeyer and Schulz-Hardt, 2003; Faulmüller et al., 2010). Perceived validity should be enhanced when arguments are shared or consistent with one's own preferences (Yaniv and Kleinberger, 2000; Volzhanin et al., 2015).

Other accounts have started to examine the cognitive basis of the shared-information bias. As shared arguments are introduced and repeated more frequently (Stasser, 1988; Stasser et al., 1989; Larson and Harmon, 2007), they have a natural memory advantage over unshared arguments. This advantage could interfere with the solution of hidden profile tasks, which draw heavily on the utilization of less well memorized unshared and preference-inconsistent items. Indeed, a number of classical studies testify to the extra persuasive impact of information repetition (Wilson and Miller, 1968; Chalmers, 1971; Cacioppo and Petty, 1979) and to the enhanced attractiveness and preference due to repeated exposure (Zajonc, 1968; Bornstein, 1989). Similar biases favoring repeated arguments can be found in a few hidden-profile studies (Van Swol et al., 2003; Schulz-Hardt et al., 2016).

However, despite the evidence on the advantage of shared or preference-consistent arguments, hidden-profile research has so far not considered an alternative explanation in terms of the simple and uncontested principle that all inductive learning increases with the number of trials. Without any group discussion or prior commitment to individual preferences, and independent of motivational factors such as social utility or subjective validity of arguments, when every item is given the same attention in an unbiased process, evaluative-learning should reflect the number of trials providing positive and negative evidence for different targets. For every stimulus item linking a target to a positive (negative) stimulus item, an increment (decrement) should be added to the evaluation of that target. This valence-updating process should be sensitive to repetitions, not only to novel stimuli, as evident from work on evaluative conditioning (Hofmann et al., 2010) and instance-based learning (Gonzalez and Dutt, 2011). Thus, an unbiased learning mechanism affords a sufficient explanation of the impact of repetition, independent of motivated biases like social sharing, preference consistency, or social validation (Boos et al., 2013).

While such an unbiased, ordinary-learning account calls for the manipulation of repetition as independent variable, almost all previous studies have treated repetition as a dependent variable, showing that shared information is likely to be repeated. Moreover, the two available publications by Van Swol et al. (2003) and by Schulz-Hardt et al. (2016) rely on restricted task set-ups (e.g., including only two-choice alternatives rather than profiles over several targets; convenient protocol sheets reducing memory demands; repetition confounded with preference consistency). Theoretically, both studies focus on distinct cognitive illusions. Van Swol et al. (2003) interpret the obtained repetition bias in terms of a truth bias (Arkes et al., 1991; Boehm, 1994). A similar point is made by Weaver et al. (2007), who argue that the enhanced fluency of repeated arguments should produce a repetition bias, regardless of social validation. Schulz-Hardt et al. (2016) believe in a projective variant of social validation, assuming that repetition leads people to infer that other people share repeated opinions.

TABLE 1 | Structure of a hidden-profile problem.

	Candidate A									**Candidate B**								
	Positive						**Negative**			**Positive**			**Negative**					
Overall	**1+**	**2+**	**3+**	**4+**	**5+**	**6+**	**1–**	**2–**	**3–**	**7+**	**8+**	**9+**	**4–**	**5–**	**6–**	**7–**	**8–**	**9–**
J1	1+			4+			1–	2–	3–	7+	8+	9+	4–			7–		
J2		2+			5+		1–	2–	3–	7+	8+	9+		5–			8–	
J3			3+			6+	1–	2–	3–	7+	8+	9+			6–			9–

Natural numbers represent positive(1+… 9+) and negative arguments (1– … 9–). Although the overall information clearly shows that Candidate A is superior to Candidate B, the information available to each of three individual judges (J1,J2, J3) raises a more positive impression of B than A. This is because A's deficits and B's assets are shared (dark gray) whereas A's assets and B's deficits are unshared (light gray).

Ordinary Learning and Metacognitive Myopia

The aim of the present article is different from all previous work on hidden profiles. Starting from basic premise that learned evaluations are sensitive to the number of trials, we provide participants with unequal opportunities to learn positive and negative evaluations of four target persons. Impression judgments should reflect the number of trials conveying targets' assets and deficits. Whether an argument is new or redundant, whether repeated arguments stem from the same or from independent sources, whether learning experience is fluent or effortful, taking place in group discussions or individual encounters, a basic prediction says: evaluation learning is an increasing monotonic function of the frequency of positive minus the negative arguments.

To be sure, amount of information may be reduced when the stimulus series involves repeated, overlapping, or fully redundant arguments. Yet, merely repeating the same stimuli benefits learning. Although novel and surprising stimuli trigger better learning (Rescorla and Wagner, 1972; Sutton and Barto, 1981), a more fundamental rule says that all trials, whether novel or repetitive, will benefit learning. Even plain repetitions foster rehearsal, elaborate encoding, and consolidation and decrease the chances that arguments will be lost, overlooked, or forgotten.[1] This basic assumption not only accounts for a variety of biases in judgment and decision making (Fiedler, 1996; Fiedler et al., 2002; Lightle et al., 2009). It also offers a new perspective on hidden profiles.

For an experimental demonstration, it is necessary to deprive the hidden-profile task of other influences but repetition. Such a modified set-up appears in **Table 2**; it is the stimulus distribution used in the experiments below. The entire profile of all information about four candidates, A, B, C, D is available to all individual participants, indicating a clear-cut preference order D > C > B > A.[2] There is no group discussion, no motive to defend one's predetermined individual preferences, and no distinction of shared and unshared information. However, the selective repetition of part of the arguments creates a conflict between actual set sizes and repetition frequencies of positive and negative attributes. Although B is clearly inferior to D, B's fewer assets are repeated more often and B's more deficits are repeated less often than Ds assets and deficits, respectively, making it easier to learn assets and harder to learn deficits in B than in D. Judgments should thus exhibit a bias to favor B over D.

In the present set-up, finding the hidden profile of substantial information requires judges to ignore (the repeated) part of the superficially presented information, unlike the common task set-up in which the hidden profile includes additional (unshared) items. Thus, our design highlights the independence of the concept "hidden profile" of the specific case involving unshared items.

Metacognitive Monitoring and Correction

Because most collective learning is subject to selective repetition – due to unequal rates of majority and minority groups and variation in the information revealed by the environment – some arguments are more likely to be presented and repeated than others. But should it really be impossible to overcome this problem?

Taking a metacognitive perspective suggests an answer and a possible remedy. Because unequal sample sizes and repetition rates are ubiquitous in the real world, *homo sapiens* should have evolved meta-cognitive devices to monitor and correct for the impact of repetition. In the hidden-profile paradigm, selective repetition ought to be detected and correct for (e.g., B should be downward-corrected and D should be upward-corrected). From such a metacognitive theory perspective, it is not sufficient to point out that ordinary learning is sensitive to repetition; it is also necessary to explain why repetition and unequal validity are not corrected for.

The present approach relates an ordinary learning account to the intriguing notion of metacognitive myopia (Fiedler, 2000, 2012). Numerous findings demonstrate that sampling biases and repetition biases remain undetected and uncorrected at the metacognitive level (Fiedler et al., 2000, 2002, 2016; Unkelbach et al., 2007; Fiedler, 2012; Powell et al., 2017). For instance, Unkelbach et al. (2007) asked participants to assess how often 10 different shares were among the daily winners in a stock-market game. On some days, they watched two TV programs so that the winners were presented twice, creating a repetition bias in favor of these repeated daily winners. The chief determinant of the resulting evaluations and share preferences was the presentation frequency, regardless of whether presentations reflected new winning outcomes or mere repetitions. Strong and robust repetition bias persisted even when participants were deliberately warned to avoid being misled by mere repetitions.

Because of many similar findings in various paradigms (for a review, see Fiedler, 2012), we expected metacognitive-myopia to extend to hidden profiles. Learned preferences should be markedly biased, due to the failure to correct for apparent repetitions. Even explicit debriefing and warnings to ignore repetitions should not eliminate the bias. This expectation is easy to understand theoretically. One cannot tell one's cognitive system to stop learning from repetitions (cf. Koriat, 1997; Fiedler et al., 2016; Powell et al., 2017), just as one cannot instruct oneself to stop learning from repeated CS-US pairings in Pavlovian conditioning.

Previous work on hidden profiles never mentioned the need for metacognitive monitoring and control, although metacognitive constructs were considered. Thus, Schulz-Hardt et al. (2016) assumed that discussion partners' repetitions will reinforce the subjective validity rather than triggering an attempt to correct for repetition bias. Similarly, Weaver et al.'s (2007) notion that fluency mediates the evaluation of repeated arguments is suggestive of naïve and uncritical influences of metacognitive cues. The notion of metacognitive myopia is fundamentally different. We argue that a comprehensive account must not only explain why repetition biases (and feelings of

[1] A corollary of this account is that memory overload can increase redundancy gains (Tindale and Sheffey, 2002).

[2] Schulz-Hardt et al. (2016) used only two options that only differed in repetitions.

fluency or social validity, and countless other biases) arise in the first place. It must also explain why repetition biases go undetected and uncorrected at the metacognitive level.

Preview of Experiments and Predictions

For an empirical test of these considerations, we exposed individual participants to an audio-recorded protocol of verbal descriptions of positive and negative attributes of four target persons (A, B, C, and D). A cover story explained that targets were applicants for flat share and that the stimulus descriptions reflected the flat mates' experiences with different subsets of applicants. To rule out group dynamics and social reward motives, participants were not engaged in group discussions but were individually exposed to a pooled (audiotaped) profile.

The four applicants varied in the *effective number* of positive versus negative attributes, such that the unequivocally correct preference order (D > C > B > A) should be apparent in a no-repetition baseline condition. However, by selectively repeating subsets of the targets' positive and negative attributes (**Table 2**), the resulting *presentation frequencies* yielded a new ordering. This should cause a shift from the correct order D > C > B > A toward the repetition-based ordering B > D > A > C in Experiment 1. We expected that judges would fail to correct for repetition spontaneously. Even an explicit warning not to be misled by repetitions in one of two conditions should not undo the basic repetition effect on evaluative learning. Experiment 2 was devoted to another aspect of meta-cognitive myopia, namely, low sensitivity to variation in social validation. A repetition bias should be obtained regardless of whether repetitions came from the same source or from different flat mates (implying social validation).

In Experiment 3, the design was extended to include recall and recognition measures in addition to evaluative ratings, to substantiate the assumption that repetition fosters learning. To increase the reliability of memory tests, the number of items was doubled and four different patterns of target-item allocations served to enhance the external validity.

Moreover, Experiment 3 allowed for a more refined test of the meta-cognitive inability to correct one's evaluative judgments. Instead of instructions not to learn from repetitions, which may be impossible, participants in one condition were informed that repetitions came from one flat mate who had vested interests in manipulating the decision. Such a cheater-detection prompt (Cosmides, 1989) entails an obvious demand to correct the final ratings of D relative to B. The vested-interest scenario should therefore motivate a local correction. However, the correction should not undo the impact of selective repetition on implicit learning, as evident in a persistent repetition bias in recall and recognition. Thus, despite the local correction of immediate ratings, the memory data may reveal that repetition biases have become an irreversible social reality.

TABLE 2 | Two stimulus distributions (Series 1 and Series 2) used to study repetition biases.

	Candidate A		**Candidate B**		**Candidate C**		**Candidate D**	
Stimulus distributions used for Experiments 1 and 2								
Series 1	+	−	+	−	+	−	+	−
Effective	2	4	3	3	4	2	6	2
Presented	5	5	7	3	4	6	6	4
Arguments selected in pretesting	*64 64*	*28*	*53 53*	*38*	*6*	*79 79 79*	*56*	*5 5*
	67 67 67	*41*	*48 48*	*76*	*17*	*35 35 35*	*49*	*80 80*
		31	*12 12 12*	*27*	*7*		*22*	
		77 77			*61*		*65*	
							70	
							10	
Series 2	+	−	+	−	+	−	+	−
Effective	2	4	3	3	4	2	6	2
Presented	5	5	7	3	4	6	6	4
Arguments selected in pretesting	*22 22*	*35*	*61 61*	*77*	*67*	*31 31 31*	*12*	*27 27*
	70 70 70	*80*	*77 77*	*38*	*17*	*28 28 28*	*8*	*76 76*
		79	*10 10 10*	*5*	*53*		*65*	
		41 41			*64*		*48*	
							56	
							49	
Stimulus distribution used for Experiment 3								
	+	−	+	−	+	−	+	−
Effective	4	8	6	6	8	4	12	4
Presented	10	10	14	6	8	12	12	8

Due to selective repetition, the resulting presented frequencies of positive (+) and negative attributes (–) diverge from the effective rates of original attributes describing four candidates A, B, C, and D, due to repetition of selected items. Arguments are represented by their pretest numbers.

EXPERIMENT 1

Methods

Participants and Design

Eighty-five participants (29 males and 56 females, mean age = 23.73, *SD* = 3.75) either received course credit or 3 Euro. One participant who did not complete the major dependent measures was excluded. The remaining 84 were randomly assigned to two instruction groups (warning vs. no warning). Another group of 15 participants received the same stimulus tape, from which all repetitions had been removed, to check on the premise that without repetitions the correct preference order (D > C > B > A) can be identified. Set sizes and numbers of positive and negative attributes per target (A, B, C, and D) varied within participants (**Table 2**).

In the absence of any effect size estimates from similar research, the number of participants required to meet a power criterion was hard to estimate. Given the rather high effect sizes obtained in Experiment 1, larger samples in Experiments 2 and 3 warranted overpowered tests, as evident from the evidence reported below.

Materials

In a pretest, 80 items describing positive (e.g., "He respects and pays heed to other people's privacy," "He always tries to preserve the harmony in the shared flat") and negative attributes (e.g., "He is not very hospitable," "He transfers a bad temper easily to his flat mates") were rated by 26 judges for valence and importance for flat sharing. Two different stimulus series were constructed, such that the attributes of the four targets (cf. **Table 2**) were balanced for valence and importance. Only Series 1 was used in Experiment 1. Repetitions involved slightly altered but semantically invariant paraphrases of original items (e.g., "It is very hard to get him to help with the housework" repeated as "Getting him to help with the housework is very hard"). All items were presented vocally by three male volunteers; repetitions of the same items always came from different voices (flat mates). As all information about each target was presented as a randomly ordered block, repetitions were maximally detectable. Block order was counterbalanced.

Procedure

The entire experiment took place in computer dialog. Participants were asked to imagine living in a flat with four people, looking out for a new flat mate to replace one who had moved out. A casting would take place, during which applicants were interviewed by three flat mates. Not all of them were present when the applicants appeared, so the decision had to rely on a combined report of all flat mates' experiences with subsets of applicants. One experimental group received an explicit warning not to be misled by repetition: "Some attributes of applicants may be stated repeatedly. Do not incorporate these repetitions in your evaluation." This warning was not provided to the other group. Afterwards, participants rated the targets on five trait dimensions covering the meaning of the stimulus attributes (agreeable, communicative, appreciative, companionable, helpful; on graphical scales anchored "not at all" and "very much"). They also provided an overall evaluation of all candidates in response to the single item "How much would you like to share your flat with applicant X?"). All ratings were provided on graphical sliding scales; ratings were linearly transformed to numerical scores from 0 to 100. The entire experiment lasted between 10 and 15 min. The materials and computer procedures can be found under the following link: https://drive.google.com/drive/folders/1atdnNdyKAcdVhbWI6X-YOgg1itGpCkIQ?usp=sharing

Results and Discussion

In accordance with the transparency norm, all empirical data are publicly available. To get access, click on *Hidden prof* on the site below: http://www.psychologie.uni-heidelberg.de/ae/crisp/studies/index.html

Average evaluation scores were computed across all five ratings. To make sure that in the absence of repetitions the stimulus attributes induced the intended ordering of targets (D > C > B > A), 30 participants provided baseline ratings of the targets in a questionnaire (using exactly the same rating scales and instructions as indicated above). Two subgroups evaluated targets described by two different versions of the stimulus series. These baseline ratings were also used to estimate the internal consistency of the five-item evaluation, which amounts to $\alpha = 0.91$ when based on ratings averaged across all 30 judges, and $\alpha = 0.76$ when the five ratings were used to discriminate between all 120 = 30 (judges) × 4 (targets) individualized targets. For convenience, we analyzed unweighted average ratings.[3]

Baseline Impressions

Means and standard deviations of the baseline evaluation scores (without repetitions) are shown in **Table 3** (top row). Evidently, the stimulus series induced more positive impressions of the two superior targets (D,C) than the two inferior targets (B,A), although the two targets within each pair received similar ratings. While the four evaluation scores should have ideally produced a linear increase from A to D, the stepped line graphs in **Figure 1** suggest that the baseline evaluations were mainly sensitive to the difference between the two superior (D,C) and the two inferior targets (A,B).

For a statistical test of the intended baseline ordering, we followed Rosenthal and Rosnow's (1985) advice to test focused hypotheses rather than standard analyses of variance, calculating a contrast score that captures a linear increase in evaluative ratings from A to D. This contrast score was the sum of each participant's mean evaluation of A, B, C, and D, weighted by the *baseline contrast* coefficients −1.5, −0.5, +0.5, +1.5, respectively. Testing this baseline contrast against zero is tantamount to testing the discriminability of actually existing target differences, independent of repetitions. This premise was indeed met. The mean contrast score was clearly positive, $M = +26.79$, $SD = 25.16$ [CI 12.86; 40.72], $t(14) = 4.12$, $d = 2.20$, $p = 0.001$.

[3]Fully equivalent results were obtained (in all pilot tests and experiments) when the five traits were weighted proportionally to their rated relevance to flat sharing (i.e., 0.228, 0.178, 0.191, 0.216, and 0.187, respectively, for agreeable, communicative, appreciative, companionable, and helpful), as determined in further pilot testing.

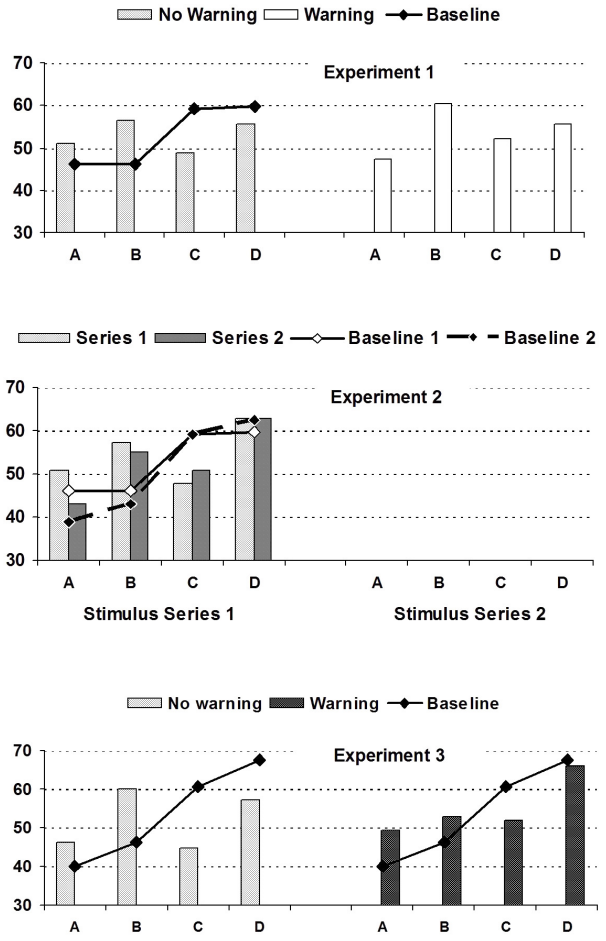


FIGURE 1 | Mean evaluative ratings (averaged across traits) of target persons A, B, C, D by experimental conditions (warning vs. no warning vs. no-repetition baseline) across experiments.

Note also that the sigmoid deviation from a purely linear trend (i.e., the slightly enhanced increase from B to C) cannot account for the repetition bias predicted for the experimental conditions (i.e., B > D > A > C), which implies that C should decrease markedly relative to B. This is evident from a *repetition contrast* defined as the sum of A, B, C, D evaluations weighted by the linear coefficients −0.5, +1.5, −1.5, +0.5, respectively, corresponding to the B > D > A > C pattern reflecting a repetition bias. Indeed, this contrast score tended to be negative, $M = -12.60$, $SD = 25.32$ [−26.62; 1.42], $t(14) = -1.93$, $d = 1.03$, $p = 0.074$, indicating that, if anything, the baseline evaluations worked against the predicted repetition bias.

A nice feature of the present design is that baseline and repetition contrasts are orthogonal; the cross product of −0.5, +1.5, −1.5, +0.5 and −1.5, −0.5, +0.5, +1.5 is exactly 0. This allows us to run independent tests of the impact of the effective number of positive and negative attributes (captured by the basic contrast) as well as the presentation frequencies (repetition contrast).

Repetition Bias on Target Evaluations

Turning to the experimental groups, the same average evaluation scores and contrast scores were used to analyze evaluations after selective repetition. As evident from the numerical means in the upper part of **Table 3** (summarized in **Figure 1**), the target evaluations reflect a mixture of both determinants, which is, however, clearly dominated by the repetition bias. Although the two superior targets C, D together received slightly higher evaluations than A, B, selective repetition caused a marked increase in the evaluation of A and B, along with a decrease in the evaluation of C and D, relative to the baseline. Explicit instructions to discount repetitions in the warning group did slightly decrease, but clearly not eliminate repetition biases (see **Figure 1**).

To disentangle the relative impact of the effective set size of different positive versus negative items and of the repetition bias, the baseline-contrast scores and the repetition-contrast scores were tested against zero. Across all 84 participants, the repetition contrast was strong and clearly above chance, $M = +14.89$, $SD = 27.99$ [+8.82; +20.96], $t(83) = 4.874$, $d = 1.064$, $p < 0.001$. The baseline

TABLE 3 | Means and standard deviations (italics) of target evaluations obtained in Experiments 1 and 2, as a function of instruction conditions (extra warning vs. no warning to ignore repetitions) and two stimulus series.

	Series 1				Series 2			
Target person	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
Baseline from pretest	46.23	46.18	59.10	59.78	38.92	43.32	59.69	62.41
	13.20	*9.46*	*13.43*	*9.96*	*12.37*	*13.20*	*11.41*	*9.85*
Experiment 1 No warning	50.99	56.42	48.75	55.41				
	12.27	*13.92*	*12.11*	*13.72*				
Experiment 1 Warning	47.36	60.43	52.16	55.43				
	11.00	*11.45*	*10.45*	*12.72*				
Experiment 2 No warning	50.96	57.09	47.76	62.81	43.16	55.01	50.85	62.86
	15.84	*13.44*	*13.93*	*10.47*	*11.70*	*13.84*	*9.80*	*11.61*

Repetitions came from different speakers in Experiment 1 but from the same speaker in Experiment 2.

contrast scores only slightly exceeded zero, $M = +5.01$, $SD = 27.26$ [−0.90; 10.93], $t(83) = 1.684$, $d = 0.367$, $p = 0.096$.

In the condition without an explicit warning to discount repeated items, only the repetition-contrast score was significantly positive $M = +13.72$, $SD = 30.44$ [4.88; 22.56], $t(47) = 3.12$, $d = 0.91$, $p = 0.003$, but not the baseline-contrast score $M = +2.79$, $SD = 27.57$ [−5.22; 10.80], $t(47) = 0.70$. This means that the repetition bias completely overrode the baseline evaluations.

An explicit warning to discount repetitions slightly increased the baseline-contrast score to a marginally significant level, $M = +7.97$, $SD = 26.94$ [−1.14; 17.08], $t(35) = 1.78$, $d = 0.60$, $p = 0.084$. However, the repetition-contrast score remained high and significant, despite the warning, $M = +16.44$, $SD = 24.70$ [8.08; 24.80], $t(35) = 3.99$, $d = 1.35$, $p < 0.001$. Indeed, the strength of repetition bias increased slightly after a warning (from 13.72 to 16.44). While this difference was far from being significant, $t(82) = 0.44$, $p = 0.661$, it highlights the ineffectiveness of the warning.

The single-item summary evaluation yielded a similar ordering as the overall evaluation score based on five trait ratings ($M = 41.26, 51.48, 45.98, 49.62$, $SD = 23.78, 25.58, 24.72, 24.72$, for A, B, C, and D, respectively). Due to the restricted reliability of this single-item measure, though, both the repetition-contrast, $M = +12.42$, $SD = 58.01$ [−0.17; 25.01], $t(83) = 1.96$, $d = 0.43$, $p = 0.053$, and the baseline contrast $M = +9.78$, $SD = 51.37$ [−1.37; 20.93], fell short of significance, $t(83) = 1.75$, $d = 0.38$, $p = 0.084$.

Altogether, these findings support the notion that even when all collective knowledge is shared, the resulting judgments are clearly biased. Mere repetitions of original items caused a marked bias in favor of A and B and against C and D, as portrayed in **Figure 1**. This finding fits a fully normal law of learning. As learning increases with repetitions, it is no wonder that the impact of repeated information on evaluations is enhanced. Yet, it is reflective of meta-cognitive myopia, the inability to correct for selective repetition.

However, as repetitions in Experiment 1 always came from different speakers, they may have been understood as social validation. Although this cannot account for the failure of explicit discounting instructions, it may have facilitated the repetition bias. To rule out this possibility, we conducted a new experiment with repetitions always coming from the same speaker. If meta-cognition is sensitive to social validation, the repetition bias should disappear, or the resulting judgments should be at least reduced relative to the different-speaker condition in Experiment 1. Conversely, if clearly redundant same-person repetitions continue to exert a systematic bias, this would lend further support to metacognitive myopia.

Another limitation of Experiment 1 was the constant assignment of attributes to targets. In Experiment 2, we used two different stimulus tapes (Series 1 and 2) with different assignments.

EXPERIMENT 2

Methods

Participants and Design

Fifty-four students (9 males and 45 females; mean age = 23.93, $SD = 6.19$) of Heidelberg University participated either for payment (3 Euro) or for course credit. The same distribution of positive and negative target attributes was used as in Experiment 1.

Materials

To rule out specific material effects, two different stimulus tapes with different assignments of specific attributes to targets (cf. **Table 2**) were assigned to different participants.

Procedure

All participants received instructions without a warning to discount repetitions. Unlike Experiment 1, all repetitions came from the same speaker, highlighted by the block-wise presentation of all items per target. Otherwise the procedure was identical to Experiment 1.

Results and Discussion

Baseline Impressions

Table 3 (right part) shows that both stimulus series led to very similar baseline evaluations, consistent with the intended ordering D > C > B > A. The mean baseline-contrast score for the new tape was highly positive, $M = 43.43$, $SD = 27.52$ [CI 28.19; 58.67], $t(14) = 6.11$, $d = 3.27$, $p < 0.001$. The mean repetition-contrast score was again negative, $M = -12.81$, $SD = 22.99$ [CI −25.54; −0.08], $t(14) = -2.16$, $d = 1.15$, $p < 0.049$. Any material bias should thus render the test of a repetition bias conservative.

Repetition Bias on Target Evaluations

Indeed, sensitivity to the effective differences in positive and negative target attributes was enhanced when repetitions came from the same speaker, thus ruling out any social-validation effect. Same-speaker repetitions apparently sensitized judges to the actual differences between targets. However, this did not eliminate or reduce the repetition bias. Though the baseline contrast score was elevated, $M = +20.30$, $SD = 28.35$ [12.56; 28.04], $t(53) = 5.26$, $d = 1.45$, $p < 0.001$, repetition contrast scores remained high and significant, $M = +18.00$, $SD = 29.14$ [10.05; 25.95], $t(53) = 4.54$, $d = 1.25$, $p < 0.001$. Both versions of the stimulus input replicated the same basic pattern (see **Table 3**).

A comparison of Experiments 1 and 2 (drawing on same materials and participant pool) corroborates the impression of similar repetition biases induced by different and same speakers, $M = +14.89$ vs. $+18.00$, $t(136) = -0.628$, $d = -0.541$, $p = 0.531$. Independent of this comparison across experiments, the strong and significant repetition bias obtained with same speakers in Experiment 2 highlights the metacognitive insensitivity to lack of social validation.

EXPERIMENT 3

So far, we have silently assumed that the tenacity of the repetition bias reflects a natural learning advantage of repeated stimuli. For an empirical check on this assumption, the design of Experiment 3 was augmented to include a free recall test and a recognition test. Both memory measures were expected to reflect the learning advantage of repeated items. To render the two memory tests sufficiently reliable, the number of stimulus items used was doubled (**Table 2**, bottom part). Thus, Experiment 3 also affords an extended replication.

While an ordinary-learning approach clearly predicts that repetition biases should be manifested in memory performance, this need not imply that the repetition effect on target evaluations is mediated by its effect on memory. It is not clear whether the final target judgments are memory based or reflective of an online process of continuous updating taking place during stimulus presentation (Hastie and Park, 1986; Hogarth and Einhorn, 1992). Such an instance-based online learning process (cf. Gonzalez and Dutt, 2011) may not produce a strong correlation between item memory and evaluative judgments. People who produce the strongest repetition bias in target ratings need not also exhibit the strongest bias in target ratings. Selective repetition (R) might be a common cause of independent biases in memory (M) and judgment (J) tasks. Experiment 3 also offers an opportunity to compare such a common-cause model $R \rightarrow M,J$ against a mediation model $R \rightarrow M \rightarrow J$.

Furthermore, Experiment 3 included a new manipulation to gain a more refined picture of the meta-cognitive inability to correct for selective repetition. Assuming that people cannot undo repetition effects on learning does not mean that they cannot correct their final judgments on demand. When told that repetitions come from flat mates with vested interests in manipulating the target evaluations, judges may easily follow the demand and downgrade B and A (who profit from repetition) relative to C and D (who suffer from repetition). But such a demand-driven correction will hardly undo the learning-advantage of repetition. The bias should still be alive in recall and recognition, waiting to become social reality and to be utilized in future judgments and communications.

Methods

Participants and Design

Ninety-five students of the University of Heidelberg (27 male, 68 female; average age 23.14, *SD* = 4.03) participated either for payment (6 Euro) or to meet a course requirement. They were randomly assigned to two experimental groups (warning vs. no warning) that only differed in whether or not instructions provided a warning to ignore selective repetitions from flat mates with vested interests. Recall and recognition tests were included along with the target ratings. A separate baseline group (n = 80) rated the four targets based on one of four new stimulus series without repetitions.

Materials and Procedures

The same materials and procedures were used as in all previous experiments, except for three distinct changes. First, the number of stimulus attributes was doubled to base the memory tests on a reasonable number of items. As shown in **Table 2**, the number of attributes was now 12 for targets A, B, and C and 16 for target D. The presentation frequencies resulting from selective repetition were also twice as high as in Experiments 1 and 2.

The selection and pre-scaling of the enlarged stimulus set were accomplished in a new pilot study, in which 28 judges rated 114 attributes relevant to flat sharing for valence. Four different versions of the stimulus series were constructed, balanced for importance of positive (e.g., "It is important for him that he has a good relationship with his roommates," "He takes care and respects the inventory in the flat") and negative behaviors associated with the four targets (e.g., "He is not very dependable"). All 80 items (52 basic attributes plus 28 repetitions; cf. **Table 2**) were tape-recorded and presented vocally by three male volunteers.

In the warning condition, all positive repetitions of target B and all negative repetitions of C and D always came from the same voice (one for each target), consistent with the suggestion that someone had vested interests in upgrading or downgrading one particular target. Repetitions of target A attributes came from all three voices. According to the instructions, target A was known by all speakers because they had recently met him at a birthday party. Target B was said to be a study mate of one speaker, who was therefore interested in B's help on home work and exam preparation. C was said to be unwanted by another speaker, because they both owned a car and they would have to compete for a single parking slot. The reason for the third speaker to avoid target D was that bathroom conflicts could be anticipated because both had to get up and rush to work early in the morning. Pragmatically, then, it was easy to see that B ratings ought to be downward-corrected whereas C and D ought to be upward-corrected.

In the no-warning condition, the stimulus series consisted of four counterbalanced blocks of target descriptions presented by the same speaker (and thereby minimizing social validation).

Two computerized memory tests were presented at the end of the session. The recall test always preceded the recognition test. Participants were asked to write down all attributes they could recall in separate text fields for targets "A," "B," "C," and "D" (presented in random order). Responses were scored as correct if they reflected the correct target reference and the substance of an original item, according to two independent coders who were blind for conditions (Cohen's Kappa = 0.92). Separate recall proportions were calculated for singular and repeated attributes (pooling double and triple repetitions).

The final 71-item recognition test consisted of all 52 original items intermixed with 19 new items that had been never presented. Items were presented on head phones in random order. Participants were then asked on screen, without time constraints, whether the prompted item had been included in the list. If the answer was "Yes," they had to indicate the target with which the item had been associated. We also assessed the confidence of recognition responses but refrained from analyzing these data. Two separate measures were

calculated, the proportion of correctly recognized unrepeated items and the corresponding recognition proportion for repeated items.

Results and Discussion

Baseline Impressions

We first of all conducted a test of the premise that, in the absence of repetitions, the effective number of targets' positive and negative attributes would produce the ordering $D > C > B > A$. The mean evaluation scores clearly increased as intended from A to D (see **Table 4** and solid lines in **Figure 1**). Closer analyses revealed that this premise held for all four versions of the stimulus tape. For an empirical check, we computed the same baseline contrast scores as in previous experiments, summing up the evaluation scores of targets A, B, C, and D weighted by the contrast coefficients −1.5, −0.5, +0.5, +1.5. The average baseline contrast score in the baseline condition was highly positive, $M = +49.89$, $SD = 30.85$ [43.03; 56.75] and different from zero, $t(79) = 14.46$, $d = 3.25$, $p < 0.001$.

Again, we also computed the repetition-contrast scores to rule out the possibility that the expected repetition bias in the experimental conditions may be peculiar to specific stimuli. Contrary to such a bias, the repetition contrast score (i.e., A, B, C, D ratings weighted by coefficients −0.5, +1.5, −1.5, +0.5) actually tended to take on a negative value, $M = -7.71$, $SD = 24.54$ [−13.17; −2.25], $t(79) = -2.81$, $d = 0.63$, $p > 0.001$. Thus, as in Experiments 1 and 2, the baseline impressions were slightly working against the experimental prediction; a contrast capturing the repetition pattern B > D > A > C tended to be negative.

Repetition Bias on Target Evaluations

Despite this conservative bias in the stimulus materials, the introduction of selective-repetitions caused a strong shift toward positive repetition-contrast scores. Across both conditions, the distribution of repetition contrast scores was clearly above zero, reflecting the expected repetition bias, $M = +18.45$, $SD = 28.50$ [12.64; 24.26], $t(94) = 6.310$, $d = 1.295$, $p < 0.001$. The baseline-contrast score was also significant across both conditions, $M = +15.77$, $SD = 34.46$ [8.75; 22.79], $t(94) = 4.461$, $d = 0.915$, $p < 0.001$.

The strength of the repetition bias, however, was moderated by the warning manipulation. When participants did not receive a warning that selective repetitions came from speakers with vested interests, the repetition-contrast score was positive and highly significant, $M = +28.34$, $SD = 25.82$ [20.76; 35.92], $t(46) = 7.60$, $d = 2.22$, $p < 0.001$. As in previous experiments, a marked repetition bias was manifested in elevated ratings of target B but lower evaluations of target D (see **Table 4** and **Figure 1**). The baseline-contrast score fell short of significance, $M = +7.90$, $SD = 32.34$ [−1.60; 17.40], $t(46) = 1.691$, $d = 0.493$, $p = 0.098$. Apparently, then, judgments were no more sensitive to independent attributes than to repetitions.

However, in the warning condition, the repetition bias was greatly reduced, though not fully eliminated. The mean repetition-contrast score, $M = +9.56$, $SD = 28.55$ [1.27; 17.85], was still positive, $t(47) = 2.35$, $d = 0.68$, $p = 0.023$, though clearly lower than in the no-warning group (cf. **Table 4**), $t(93) = 3.39$, $d = 0.70$, $p = 0.001$.[4]

The effect of warning conditions testifies to judges' ability to modify their ratings in accordance with explicit hints to deceptive behavior (bottom chart of **Figure 1**). At the same time, the blatant warning served to strengthen the original (baseline) ordering, as evident in positive and significant baseline-contrast scores, $M = +24.34$, $SD = 35.04$ [14.16; 34.52], $t(47) = 4.86$, $d = 1.40$, $p < 0.001$, which were higher than the baseline contrast scores in the no-warning condition, $M = 7.90$, $SD = 32.34$, $t(93) = 2.227$, $d = 0.459$, $p = 0.028$.

Apparently, then, when participants know that selective repetitions serve a manipulative goal, they are capable of correcting their final ratings. If speakers have vested interests in benefitting B and harming C and D, judges know how to correct for the bias: one only has to downgrade B ratings and to upgrade C and D ratings. The crucial question, though, is whether this correction eliminates the mental extract of the repetition bias or whether it merely changes the overt judgment output. The correction might remain superficial while the repetition bias might live on in the judges' memory, waiting to influence later judgments or actions. Both memory measures afford a straightforward test of this challenging issue. Even though participants were apparently able to correct for a bias on overt rating scales, they may not be able to undo the uncontrollable effect of stimulus repetition on recall and recognition.

Recall

Indeed, across all participants, the correct-recall proportion of repeated items was much higher, $M = 0.120$, $SD = 0.100$ [CI 0.109; 0.150], than proportions of recalled unrepeated items, $M = 0.059$, $SD = 0.054$ [CI 0.048; 0.070], $t(94) = 6.83$, $d = 1.39$, $p < 0.001$. This recall advantage of repeated items was similarly strong in the no-warning condition, $M = 0.126$, $SD = 0.104$ [CI 0.096; 0.156], versus $M = 0.064$, $SD = 0.057$ [CI 0.047; 0.081], $t(47) = 4.95$, $d = 1.44$, $p < 0.001$, as in the warning condition, $M = 0.113$, $SD = 0.097$ [CI 0.085; 0.141], versus $M = 0.054$, $SD = 0.051$ [CI 0.039; 0.069], $t(48) = 4.66$, $d = 1.35$, $p < 0.001$. Thus, the blatant warning did not reduce the strength of the repetition bias in recall, regardless of the corrections applied to the immediate target ratings.

Recognition

The analysis of the recognition data provided further support for the persistence of the repetition bias, although the pattern was not quite the same. Across all participants, responses on the combined recognition and assignment test were more likely to be correct for repeated items, $M = 0.372$, $SD = 0.145$ [CI 0.342; 0.401], than for unrepeated items, $M = 0.324$, $SD = 0.102$ [CI 0.302; 0.346], $t(94) = 2.91$, $d = 0.60$, $p = 0.005$. However, notably, this tendency was not significant in the no-warning condition, $M = 0.361$, $SD = 0.155$ [CI 0.315; 0.406],

[4]The only significant result in a 2 (warning conditions) × 4 (material versions) ANOVA was a main effect for conditions, $F(1,87) = 11.181$, $p = 0.001$. Neither the material versions main effect, $F(3,87) = 1.775$, $p = 0.158$, nor the interaction was significant, $F(3,87) = 0.402$, $p = 0.752$, reflecting a robust effect that is not peculiar to specific stimuli.

versus $M = 0.327$, $SD = 0.100$ [CI 0.298; 0.356], $t(46) = 1.53$, $d = 0.45$, $p = 0.132$. Ironically, it was stronger in the warning condition, $M = 0.382$, $SD = 0.135$ [CI 0.343; 0.422], versus $M = 0.321$, $SD = 0.115$ [CI 0.288; 0.354], $t(47) = 2.52$, $d = 0.74$, $p = 0.015$.

Thus, although the repetition bias was somewhat weaker in recognition than in recall, the evidence from both memory tests supports the notion that the bias persisted in memory although deliberate responses on rating scales could be corrected in accordance with instruction demands. Even when stimulus input is strongly discredited, metacognition can hardly tell the cognitive system not to learn from repetition. An interesting question for future research is whether the memory advantage of repeated information also persists after a longer delay.

Relating Judgment Biases to Memory Measures

Finally, it is interesting to examine the relationship between individual judges' repetition contrast scores and their corresponding biases in the two memory tasks. In fact, both correlations turned out to be low. Individual differences in the repetition contrast scores were only weakly correlated with individual measures of the differential proportions of correctly recalled repeated items minus correctly recalled singular items, $r(df = 93) = 0.152$, $p = 0.142$. When computed separately per condition, this correlation was close to zero without a warning, $r(df = 45) = 0.063$, $p = 0.676$, and slightly higher after a warning, $r(df = 46) = 0.230$, $p = 0.116$. The corresponding correlations between repetition contrast scores and differential recognition proportions for repeated minus singular items were negligible: $r(df = 93) = -0.051$, $p = 0.625$ across all participants; $r(df = 45) = -0.016$, $p = 0.917$ without a warning, and $r(df = 46) = -0.029$, $p = 0.845$ after a warning.

Relying on a total of 95 participants, these small correlations can be hardly attributed to insufficient statistical power. Although the present experiments were not designed to allow for strict tests of the underlying mechanism, the range of correlations is hardly compatible with the assumption that evaluative biases are substantially mediated by selective memory biases. Much more likely than a memory-based judgment process is the assumption that evaluations are learned online (Hastie and Park, 1986) and that evaluative ratings and memory responses are influenced by the same common cause.

GENERAL DISCUSSION

The hidden-profile paradigm continues to fascinate scientists; it is at the heart of democratic culture. Democracies deputize decisions to collectives relying on a division of labor, calling for the integration of the knowledge and expertise of several agents or advisors. However, the available evidence (Kerr and Tindale, 2004) shows that people have a hard time to coordinate and exploit collective knowledge. Three decades of illuminating experimental research in the hidden-profile paradigm testify to this problem.

The failure to solve hidden profiles has been explained in terms of such group-dynamic factors as the reward and the social validation value of shared information (Wittenbaum et al., 1999; Greitemeyer and Schulz-Hardt, 2003), the memory advantage of shared over unshared arguments (Lightle et al., 2009), and decision schemes favoring arguments consistent with pre-existing individual preferences (Edwards and Smith, 1996; Schulz-Hardt et al., 2016). Prior research has also noted that shared and preference-consistent arguments are likely to be repeated and that repetition (Schulz-Hardt et al., 2006; Stasser et al., 2012) and resulting feelings of fluency (Weaver et al., 2007) can influence subsequent target judgments.

However, prior research and theorizing never elaborated on the fundamental rule that all evaluative learning increases with the number of trials and that frequency of presentation and repetition are therefore primary causal variables that can explain troubles with hidden profiles independently of motivated biases and group-dynamics. Even when all items are shared and when there is no extra motive to process particular items more than others, the presentation rate of different information items will always vary as a function of many environmental conditions. In the absence of any bias to attend to or to elaborate on specific arguments more than on others, presentation rates will be higher for majority than minority arguments, for proximal than distal events, for ingroups than outgroups, and for public than private knowledge, to list but a few ecological determinants of item frequency.

From such an ordinary-learning perspective, information sharing, preference consistency, and social validation are only special cases of a much broader class of environmental causes of selective presentation and repetition. Even when all the

TABLE 4 | Means and standard deviations (italics) of target evaluations obtained in Experiments 3 as a function of instruction conditions (warning vs. no warning).

	No warning				Warning			
Target person	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
Baseline (no repetition)	39.93	46.17	60.59	67.59				
	12.99	*12.11*	*11.05*	*11.87*				
Version 1	43.07	50.20	62.99	65.73				
Version 2	35.37	50.28	58.50	69.47	Same baseline data hold for the no-warning and the warning condition			
Version 3	39.26	39.72	57.57	65.44				
Version 4	39.84	42.16	61.96	70.64				
Target evaluations	46.39	60.03	44.78	57.31	49.48	52.76	51.88	65.99
	13.21	*14.20*	*12.39*	*12.62*	*13.22*	*15.36*	*12.07*	*14.24*

prominent factors emphasized in previous research are controlled for, one cannot expect the problem with hidden profiles to be resolved because other, often quite normal, factors will continue to create selective repetition. Some daily news are encountered twice or more often; sometimes "breaking news" is indeed recycled abundantly; the same emails often reach us multiple times; misunderstandings or debates may motivate argument repetitions, or in democratic discussions, the same points are stated more frequently if held by a majority (Fiedler and Wänke, 2009). And last not least, in science, frequently cited findings have a clear-cut repetition advantage.

Because repetition biases are ubiquitous and because it is impossible not to learn from repetition, a major role is assigned to metacognition. If repetition biases are unavoidable in the first place, because the opportunity to learn is never the same across all items, monitoring and control functions are required to detect and correct for repetition biases. However, the present findings demonstrate that participants fail to correct for selective repetition. Although the blocked presentation mode facilitated the detection of repetitions, and when repetitions coming by the same speaker minimized their social validation value, evaluative judgments continued to be biased toward selective repetitions. Moreover, even when judges were sensitized through explicit debriefing and instructions not to be misled by selective repetitions, ruling out pragmatic demands to consider repeated arguments valid, the repetition bias could not be erased.

Note that the causal role assigned to metacognitive monitoring and control is independent of whether repetition rates might be correlated with other factors such as fluency, inferred consensus, or subjective validity. The theoretical importance of metacognition is independent of such natural confounds of repetition frequency. Regardless of what experiential cues drive the enhanced impact of repeated arguments – fluency, social validity, or sharedness – the correction of unavoidable repetition biases calls for metacognitive monitoring and control functions.

It is not too surprising, of course, that metacognitive control cannot undo automatic learning. One cannot tell the cognitive system to cease learning from repetition, just as we cannot tell our body to cease learning from repeated pairing of conditional and unconditional stimuli in Pavlovian conditioning. Unsurprising as this contention may be, it has distinct and memorable implications for collective judgments and decisions in democratic societies. Simply allowing every argument to be presented is no guarantee for unbiased judgments and decisions. Rarely presented minority arguments will more likely be ignored, forgotten or overridden than frequently presented majority arguments. To correct for this inevitable learning asymmetry, it would be necessary to allow rare arguments or minority positions to be presented more often than common majority positions. However, such an ironic minority privilege would not be compatible with the spirit of democracy either. Democratic rules alone cannot solve the dilemma. Rather, the burden of rational decision making rests on democratic agents' meta-cognitive ability to distinguish valid from invalid, original arguments from redundant repetitions.

The present findings strongly suggest that metacognitive myopia prevents homo sapiens from this kind of critical assessment, adding convergent evidence to existing findings on metacognitive myopia (Fiedler, 2000, 2012; Fiedler et al., 2016).

We anticipate that it will hardly be possible to prevent the initial occurrence of repetition biases in the first place. We rather believe that the existence of this ubiquitous source of bias must be taken for granted as a natural product of environmental learning. It can only be diagnosed and corrected at the metacognitive level. However, a host of convergent evidence suggests that metacognitive myopia prevents homo sapiens from critical assessment and correction (Fiedler, 2000, 2012; Fiedler et al., 2016). Even explicit reminders not to be misled by selective repetition and lopsided sampling do not prevent people from adopting the more frequently presented arguments. The present findings corroborate this conclusion in the context of collective judgments: even when social validation is ruled out and when blocked presentation makes argument repetition maximally visible, and sometimes even after a warning to avoid a repetition bias, participants continue to be strongly influenced by mere repetition.

Note that metacognitive myopia affords a functionalist account rather than a mechanistic (Fiedler, 2016). It highlights the failure to engage in metacognitive monitoring and control functions, which might involve a variety of different mental algorithms. For some reason, *homo sapiens* is not sufficiently motivated or may have actively learned not to engage in retrograde correction of even blatant sampling biases (Fiedler, 2008, 2012). We exhibit perseverance after full debriefing that some feedback was completely wrong (Ross et al., 1975); we continue to be influenced by fake news after debunking (Chan et al., 2017), we treat advertising as a source of evidence and citation rates as a symptom of good science, without any attempt to control for obvious sampling biases.

One may speculate that metacognitive myopia serves adaptive functions, conserving one's faith in the validity of the empirical world and preventing people from tedious correction processes for which there is often no normative solution. Alternatively, there may have been insufficient selection pressure, maybe because metacognitive monitoring and control has only lately become important during a rather short information era, or it may simply not constitute a genuine survival advantage. Nevertheless, in the context of specific problems, such as personnel selection or investment decisions, it would be beneficial to develop decision aids and training programs to overcome the constraints of metacognitive myopia, to avoid injustice and irrational action.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the German Association of Psychology (DGPs). All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The research underlying the present article was supported by a DFG grant (Fi 293/26-1) awarded to KF.

REFERENCES

Arkes, H. R., Boehm, L. E., and Xu, G. (1991). Determinants of judged validity. *J. Exp. Soc. Psychol.* 27, 576–605. doi: 10.1016/0022-1031(91)90026-3

Boehm, L. E. (1994). The validity effect: a search for mediating variables. *Pers. Soc. Psychol. Bull.* 20, 285–293. doi: 10.1177/0146167294203006

Boos, M., Schauenburg, B., Strack, M., and Belz, M. (2013). Social validation of shared and nonvalidation of unshared information in group discussions. *Small Group Res.* 44, 257–271. doi: 10.1177/1046496413484068

Bornstein, R. F. (1989). Exposure and affect: overview and meta-analysis of research, 1967– 1987. *Psychol. Bull.* 106, 265–289. doi: 10.1037/0033-2909.106.2.265

Brauer, M., Judd, C. M., and Gliner, M. D. (1995). The effects of repeated expressions on attitude polarization during group discussions. *J. Pers. Soc. Psychol.* 68, 1014–1029. doi: 10.1037/0022-3514.68.6.1014

Cacioppo, J. T., and Petty, R. E. (1979). Effects of message repetition and position on cognitive response, recall, and persuasion. *J. Pers. Soc. Psychol.* 37, 97–109. doi: 10.1037//0022-3514.37.1.97

Chalmers, D. K. (1971). Repetition and order effects in attitude formation. *J. Pers. Soc. Psychol.* 17, 219–228. doi: 10.1037/h0030379

Chan, M. S., Jones, C. R., Hall Jamieson, K., and Albarracín, D. (2017). Debunking: a meta-analysis of the psychological efficacy of messages countering misinformation. *Psychol. Sci.* 28, 1531–1546. doi: 10.1177/0956797617714579

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

Dennis, A. R. (1996). Information exchange and use in small group decision making. *Small Group Res.* 27, 532–550. doi: 10.1177/1046496496274003

Diehl, M., and Stroebe, W. (1987). Productivity loss in brainstorming groups: toward the solution of a riddle. *J. Pers. Soc. Psychol.* 53, 497–509. doi: 10.1037/0022-3514.53.3.497

Edwards, K., and Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *J. Pers. Soc. Psychol.* 71, 5–24. doi: 10.1037/0022-3514.71.1.5

Faulmüller, N., Kerschreiter, R., Mojzisch, A., and Schulz-Hardt, S. (2010). Beyond group-level explanations for the failure of groups to solve hidden profiles: the individual preference effect revisited. *Group Process. Intergroup Relat.* 13, 653–671. doi: 10.1177/1368430210369143

Faulmüller, N., Mojzisch, A., Kerschreiter, R., and Schulz-Hardt, S. (2012). Do you want to convince me or to be understood?: preference-consistent information sharing and its motivational determinants. *Pers. Soc. Psychol. Bull.* 38, 1684–1696. doi: 10.1177/0146167212458707

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. *Psychol. Rev.* 103, 193–214. doi: 10.1037/0033-295X.103.1.193

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. (2008). The ultimate sampling dilemma in experience-based decision making. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 186–203. doi: 10.1037/0278-7393.34.1.186

Fiedler, K. (2012). "Meta-cognitive myopia and the dilemmas of inductive-statistical inference," in *The Psychology of Learning and Motivation*, Vol. 57, ed. B. H. Ross (San Diego, CA: Elsevier Academic Press), 1–55.

Fiedler, K. (2016). Functional research and cognitive-process research in behavioural science: an unequal but firmly connected pair. *Int. J. Psychol.* 51, 64–71. doi: 10.1002/ijop.12163

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

Fiedler, K., Kareev, Y., Avrahami, J., Beier, S., Kutzner, F., and Hütter, M. (2016). Anomalies in the detection of change: when changes in sample size are mistaken for changes in proportions. *Mem. Cogn.* 44, 143–161. doi: 10.3758/s13421-015-0537-z

Fiedler, K., Walther, E., Freytag, P., and Plessner, H. (2002). Judgment biases in a simulated classroom – a cognitive-environmental approach. *Organ. Behav. Hum. Decis. Process.* 88, 527–561. doi: 10.1006/obhd.2001.2981

Fiedler, K., and Wänke, M. (2009). The cognitive-ecological approach to rationality in social psychology. *Soc. Cogn.* 27, 699–732. doi: 10.1521/soco.2009.27.5.699

Gonzalez, C., and Dutt, V. (2011). Instance-based learning: integrating decisions from experience in sampling and repeated choice paradigms. *Psychol. Rev.* 118, 523–551. doi: 10.1037/a0024558

Greitemeyer, T., and Schulz-Hardt, S. (2003). Preference-consistent evaluation of information in the hidden profile paradigm: beyond group-level explanations for the dominance of shared information in group decisions. *J. Pers. Soc. Psychol.* 84, 322–339. doi: 10.1037/0022-3514.84.2.322

Hastie, R., and Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychol. Rev.* 93, 258–268. doi: 10.1037/0033-295X.93.3.258

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., and Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychol. Bull.* 136, 390–421. doi: 10.1037/a0018916

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

Kerr, N. L., and Tindale, R. S. (2004). Group performance and decision making. *Annu. Rev. Psychol.* 55, 623–655. doi: 10.1146/annurev.psych.55.090902.142009

Koriat, A. (1997). Monitoring one's own knowledge during study: a cue utilization approach to judgments of learning. *J. Exp. Psychol. Gen.* 126, 349–370. doi: 10.1037/0096-3445.126.4.349

Larson, J. J., and Harmon, V. M. (2007). Recalling shared vs. unshared information mentioned during group discussion: toward understanding differential repetition rates. *Group Process. Intergroup Relat.* 10, 311–322. doi: 10.1177/1368430207078692

Larson, J. R. Jr., Foster-Fishman, P. G., and Keys, C. B. (1994). Discussion of shared and unshared information in decision making groups. *J. Pers. Soc. Psychol.* 67, 446–461. doi: 10.1037/0022-3514.67.3.446

Lightle, J., Kagel, J., and Arkes, H. (2009). Information exchange in group decision making: the hidden profile problem reconsidered. *Manag. Sci.* 55, 568–581. doi: 10.1287/mnsc.1080.0975

Lu, L., Yuan, Y. C., and McLeod, P. L. (2012). Twenty-five years of hidden profiles in group decision making: a meta-analysis. *Pers. Soc. Psychol. Rev.* 16, 54–75. doi: 10.1177/1088868311417243

Mannes, A. E., Soll, J. B., and Larrick, R. P. (2014). The wisdom of select crowds. *J. Pers. Soc. Psychol.* 107, 276–299. doi: 10.1037/a0036677

McCauley, C. (1998). Groupthink dynamics in Janis's theory of groupthink: backward and forward. *Organ. Behav. Hum. Decis. Process.* 73, 142–162. doi: 10.1006/obhd.1998.2759

Mesmer-Magnus, J., and DeChurch, L. (2009). Information sharing and team performance: a meta-analysis. *J. Appl. Psychol.* 94, 535–546. doi: 10.1037/a0013773

Powell, D., Yu, J., DeWolf, M., and Holyoak, K. J. (2017). The love of large numbers: a popularity bias in consumer choice. *Psychol. Sci.* 28, 1432–1442. doi: 10.1177/0956797617711291

Rescorla, R. A., and Wagner, A. R. (1972). "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory*, eds A. H. Black and W. F. Prokasy (New York, NY: Appleton-Century-Crofts), 64–99.

Rosenthal, R., and Rosnow, R. L. (1985). *Contrast Analysis: Focused Comparisons in the Analysis of Variance*. Cambridge: Cambridge University Press.

Ross, L., Lepper, M. R., and Hubbard, M. (1975). Perseverance in self-perception and social perception: biased attributional processes in the debriefing paradigm. *J. Pers. Soc. Psychol.* 32, 880–892. doi: 10.1037//0022-3514.32.5.880

Schulz-Hardt, S., Brodbeck, F., Mojzisch, A., Kerschreiter, R., and Frey, D. (2006). Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *J. Pers. Soc. Psychol.* 91, 1080–1093. doi: 10.1037/0022-3514.91.6.1080

Schulz-Hardt, S., Giersiepen, A., and Mojzisch, A. (2016). Preference-consistent information repetitions during discussion: do they affect subsequent judgments and decisions? *J. Exp. Soc. Psychol.* 64, 41–49. doi: 10.1016/j.jesp.2016.01.009

Schulz-Hardt, S., and Mojzisch, A. (2012). How to achieve synergy in group decision making: lessons to be learned from the hidden profile paradigm. *Eur. Rev. Soc. Psychol.* 23, 305–343. doi: 10.1080/10463283.2012.744440

Stasser, G. (1988). Computer simulation as a research tool: the DISCUSS model of group decision making. *J. Exp. Soc. Psychol.* 24, 393–422. doi: 10.1016/0022-1031(88)90028-5

Stasser, G., Abele, S., and Parsons, S. V. (2012). Information flow and influence in collective choice. *Group Process. Intergroup Relat.* 15, 619–635. doi: 10.1177/1368430212453631

Stasser, G., Taylor, L. A., and Hanna, C. (1989). Information sampling in structured and unstructured discussions of three- and six-person groups. *J. Pers. Soc. Psychol.* 57, 67–78. doi: 10.1037/0022-3514.57.1.67

Stasser, G., and Titus, W. (1985). Pooling of unshared information in group decision making: biased information sampling during discussion. *J. Pers. Soc. Psychol.* 48, 1467–1478. doi: 10.1037/0022-3514.48.6.1467

Stasser, G., and Titus, W. (2003). Hidden profiles: a brief history. *Psychol. Inq.* 14, 304–313. doi: 10.1080/1047840X.2003.9682897

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. New York, NY: Doubleday & Co.

Sutton, R. S., and Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.* 88, 135–170. doi: 10.1037/0033-295X.88.2.135

Tindale, R., and Sheffey, S. (2002). Shared information, cognitive load, and group memory. *Group Process. Intergroup Relat.* 5, 5–18. doi: 10.1177/1368430202005001535

Unkelbach, C., Fiedler, K., and Freytag, P. (2007). Information repetition in evaluative judgments: easy to monitor, hard to control. *Organ. Behav. Hum. Decis. Process.* 103, 37–52. doi: 10.1016/j.obhdp.2006.12.002

Van Swol, L. M., Savadori, L., and Sniezek, J. A. (2003). Factors that may affect the difficulty of uncovering hidden profiles. *Group Process. Intergroup Relat.* 6, 285–304. doi: 10.1177/13684302030063005

Volzhanin, I., Hahn, U., Jönsson, M. L., and Olsson, E. J. (2015). "Individual belief revision dynamics in a group context," in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Pasadena, CA, 2505–2510.

Weaver, K., Garcia, S. M., Schwarz, N., and Miller, D. T. (2007). Inferring the popularity of an opinion from its familiarity: a repetitive voice can sound like a chorus. *J. Pers. Soc. Psychol.* 92, 821–833. doi: 10.1037/0022-3514.92.5.821

Wilson, W., and Miller, H. (1968). Repetition, order of presentation, and timing of arguments and measure as determinants of opinion change. *J. Pers. Soc. Psychol.* 9, 184–188. doi: 10.1037/h0021251

Wittenbaum, G. M., Hubbell, A. P., and Zuckerman, C. (1999). Mutual enhancement: toward an understanding of the collective preference for shared information. *J. Pers. Soc. Psychol.* 77, 967–978. doi: 10.1037/0022-3514.77.5.967

Yaniv, I., Choshen-Hillel, S., and Milyavsky, M. (2009). Spurious consensus and opinion revision: why might people be more confident in their less accurate judgments? *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 558–563. doi: 10.1037/a0014589

Yaniv, I., and Kleinberger, E. (2000). Advice taking in decision making: egocentric discounting and reputation formation. *Organ. Behav. Hum. Decis. Process.* 84, 260–281. doi: 10.1006/obhd.2000.2909

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *J. Pers. Soc. Psychol.* 9(Pt 2), 1–27. doi: 10.1080/02699931.2010.497409

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 September 2018
doi: 10.3389/fpsyg.2018.01479

The Psychology of Uncertainty and Three-Valued Truth Tables

Jean Baratgin[1,2]*, *Guy Politzer*[2], *David E. Over*[3] *and Tatsuji Takahashi*[4]

[1] CHArt (P-A-R-I-S), Université Paris 8 & EPHE, Paris, France, [2] Institut Jean Nicod, École Normale Supérieure, Paris, France, [3] Psychology Department, Durham University, Durham, United Kingdom, [4] School of Science and Engineering, Tokyo Denki University, Tokyo, Japan

Psychological research on people's understanding of natural language connectives has traditionally used truth table tasks, in which participants evaluate the truth or falsity of a compound sentence given the truth or falsity of its components in the framework of propositional logic. One perplexing result concerned the indicative conditional *if A then C* which was often evaluated as true when *A* and *C* are true, false when *A* is true and *C* is false but irrelevant" (devoid of value) when *A* is false (whatever the value of *C*). This was called the "psychological defective table of the conditional." Here we show that far from being anomalous the "defective" table pattern reveals a coherent semantics for the basic connectives of natural language in a trivalent framework. This was done by establishing participants' truth tables for negation, conjunction, disjunction, conditional, and biconditional, when they were presented with statements that could be certainly true, certainly false, or neither. We review systems of three-valued tables from logic, linguistics, foundations of quantum mechanics, philosophical logic, and artificial intelligence, to see whether one of these systems adequately describes people's interpretations of natural language connectives. We find that de Finetti's (1936/1995) three-valued system is the best approximation to participants' truth tables.

Keywords: natural language connectives, three-valued truth tables, uncertainty, de Finetti's tri-event, subjective probability

OPEN ACCESS

Edited by:
David R. Mandel,
Defence Research and Development Canada, Canada

Reviewed by:
Igor Douven,
Université Paris-Sorbonne, France
Keith Kirk Niall,
Retired, Toronto, ON, Canada

***Correspondence:**
Jean Baratgin
jean.baratgin@univ-paris8.fr

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

Received: *07 May 2018*
Accepted: *26 July 2018*
Published: *04 September 2018*

Citation:
Baratgin J, Politzer G, Over DE and Takahashi T (2018) The Psychology of Uncertainty and Three-Valued Truth Tables. Front. Psychol. 9:1479. doi: 10.3389/fpsyg.2018.01479

INTRODUCTION: THE BAYESIAN APPROACH TO THE PSYCHOLOGY OF REASONING

From the beginning of their investigations, and for nearly a century, psychologists studying human deductive reasoning considered bi-valued logic as the sole frame of reference. Their early inspiration was limited to Aristotelian syllogistic (Binet, 1902; James, 1908) but in the 1950s Piaget adopted propositional logic which he assumed to be the basis of adults' cognitive functioning (Inhelder and Piaget, 1958). The elementary connectives of natural language for negation, conjunction and disjunction were identified with the logical connectives $\neg$, $\wedge$, and $\vee$, respectively, and the indicative conditional *if A* (antecedent), *then C* (consequent) was identified with the material conditional (or implication $A \supset C$). However, in 1966, Wason observed that when people are required to make judgments about conditionals in terms of true and false, they often produce a table that differs from the material conditional. Participants consider that a conditional *if A then C* is made "true" by the *A and C* state of affairs and made "false" by the *A and not-C* state, but that the *not-A* cases (*not-A and C* and *not-A and not-C*) are "irrelevant" to the truth value of *if A then C*. Psychologists came to call this truth table "defective" to underscore participants' apparent

imperfect comprehension of the material conditional which was assumed to be the meaning of *if … then*. This "defective" conditional is represented in **Table 1** (column 1) as $C|_{"d"}A$. Wason's (1966) observation was confirmed by early experimental studies in which part of the participants required to choose or construct the states of affairs that make the sentence true or false disregard the *not-A* states (Evans, 1972) or choose the option *irrelevant* when it is offered to them (Johnson-Laird and Tagart, 1969), or spontaneously express the irrelevance of these cases (Delval and Riviére, 1975; Politzer, 1981). Whatever this table is called, it should be contrasted with the truth table for the material conditional which is true in the *not-A* cases (see **Table 1**, column 2). In addition, a "defective" biconditional (denoted by $C||_{"d"}A$ in **Table 1**, column 3) has also been observed (Delval and Riviére, 1975). It is made true by the *A and C* state of affairs, and made false by the *A and not-C* and *not-A and C* states, with the *not-A and not-C* state alone "irrelevant" (see Evans and Over, 2004, for further research on the "defective" conditional and biconditional truth tables in psychology).

Until the end of the century, the major part of the theoretical debate on deduction in cognitive psychology revolved around the format of representation of the connectives. For one stream of research the representation was assumed to be syntactic and deduction rule-governed (Rips, 1994; Braine and O'Brien, 1998) whereas for another stream it was assumed to be semantic and deduction model-based (Johnson-Laird and Byrne, 1991). Whatever the option may be, the frame of reference was still two-valued logic and the explanation of the defective table was a major item on the agenda. However, in recent years, this old model of reference has been questioned and a new approach using a probabilistic frame of reference has emerged. This new paradigm in the psychology of reasoning emphasizes that most human inferences take place when there is some degree of uncertainty about the subject matter (Oaksford and Chater, 2007, 2009; Over, 2009, 2016; Pfeifer and Kleiter, 2010; Evans, 2012; Elqayam and Over, 2013; Pfeifer, 2013; Baratgin et al., 2015; Baratgin and Politzer, 2016; Over and Baratgin, 2017; Over and Cruz, 2018). This uncertainty is found in both everyday thought and scientific inference, when people are trying to decide what they will most enjoy on a lunch menu, or to infer what has caused an outbreak of food poisoning.

This new Bayesian approach to the psychology of reasoning has received great impetus from two sets of experimental findings. The first finding is that, as claimed by the theory, people generally judge the probability of the indicative conditional, *P(if A then C)*, to be the conditional probability of *C* given *A*, *P(C|A)* (for early data see: Evans et al., 2003, 2007; Oberauer and Wilhelm, 2003, but see also Douven and Verbrugge, 2010; Vidal and Baratgin, 2017). The second finding is that participants' assessments of the conclusions of explicit deductive inferences made under uncertainty tend to be in the *coherence intervals* determined by the probability of the premises, that is, participants tend to conform to the laws of probability (Pfeifer and Kleiter, 2009, 2010, 2011; Pfeifer, 2014; Singmann et al., 2014; Cruz et al., 2015; Evans et al., 2015; Politzer and Baratgin, 2016).

An important target of the new Bayesian paradigm concerns the "defective" table mentioned earlier. Supporters of the new paradigm consider that far from being anomalous it reveals a semantics that differs from the material conditional (Baratgin et al., 2013, 2014)[1]. This point will be developed below and generalized to the basic connectives of natural language (negation, conjunction, disjunction) and also to the biconditional.

THE DE FINETTIAN APPROACH

Is there a normative framework for unifying all these experimental results? We have argued (Baratgin, 2015; Baratgin and Politzer, 2016; Over and Baratgin, 2017; Over and Cruz, 2018) that de Finetti's Bayesian subjective theory offers just such a framework. De Finetti is one of the founding fathers of modern probability theory, and the most prominent representative of subjective Bayesianism. His overall approach to probability (de Finetti, 1974) has deep psychological relevance (see Baratgin and Politzer, 2006, 2007; Baratgin, 2015, for a discussion in the field of the psychology of probability judgment). His conception of probability as subjective degree of belief, and of the assessment of probability through the well-known betting procedure, are rooted in psychological reflection.

de Finetti (1980) proposed three levels of knowledge of an event. The *objective* level, *Level 0*, corresponds to binary logic, in which every statement that expresses the occurrence or the non-occurrence of an event is objectively true or false. This is the level of events that are known for sure. It is this level that was traditionally studied in the psychology literature of reasoning, even though it is severely limited for a psychological approach, for people often do not know for sure what is true and what is false. It is also, ironically, the level of which de Finetti (2006, p. 113) says that it is "sterile" because logic has no other use than order, enumerate, and expound what is already known. A purely logical science cannot be concerned in forecasting. Hence the need to substitute this "rigid logic" with a "logic of the probable" that is the logic of everyday allowing to make predictions with regard to uncertain knowledge (de Finetti, 1977/1993, p. 494).

Beyond Level 0, de Finetti (1980) considered two other levels that are *subjective*. On *Level 1*, the event (or statement) concerns a specific object defined by its own characteristics known to the individual. An event is always conditioned on the individual's personal state of knowledge. The statements can be classified as having one of three values: *true*, characterizing an expected event that has happened; *false*, characterizing an expected event that has not happened; and *uncertain*. The value *uncertain* is to be understood as follows. It represents the subjective point of view of an individual who is wondering whether or not an event will happen or, equivalently, whether the statement that expresses the occurrence of the event is true or false. The third value reflects a

[1]Several philosophers have proposed an identical 2 × 2 "defective" table in their analysis of "if" in ordinary language, with different interpretations of the third value. According to Quine (1950) a conditional affirmation with a false antecedent is as if it had never been made. O'Connor (1951) defines a table with a third undetermined value. Dummet (1958/1959) presents a similar table where the third value corresponds to neither true nor false. Kneale and Kneale (1962, p. 135–136) suggest a "defective" table (in the sense of *defective truth function*) where the value I (denoted by "–") characterizes a truth-value gap.

TABLE 1 | The different truth tables for the conditional if A then C: two-valued (columns 1–3′) and three-valued (columns 4–7).

		1	**1′**	**2**	**3**	**3′**			**4**	**5**	**6**	**7**
A	C	$C\|_{''d''}A$	$C\|_{Fi}A$	$A \supset C$	$C\|\|_{''d''}A$	$C\|\|_{Fi}A$	A	C	$C\|_{?}A$	$C\|_{Fi}A$	$C\|_{Fa}A$	$C\|_{C}A$
T	T	T	T	T	T	T	T	T	T	T	T	T
							T	∅	?	∅	∅	∅
T	F	F	F	F	F	F	T	F	F	F	F	F
							∅	T	?	∅	∅	T
							∅	∅	?	∅	∅	∅
							∅	F	?	∅	F	F
F	T	I	∅	T	F	F	F	T	∅	∅	∅	∅
							F	∅	?	∅	∅	∅
F	F	I	∅	T	I	∅	F	F	∅	∅	∅	∅

T, true; F, false; I, Irrelevant; ∅, third value; ?, T or F or ∅;
1. $(C|_{''d''}A)$ *, the* 2 × 2 *"defective" conditional table;*
1′. $(C|_{Fi}A)$*, the 2 × 2 Finettian interpretation of 1;*
2.$(A \supset C)$*, the 2 × 2 material conditional;*
3. $(C||_{''d''}A)$ *, the* 2 × 2 *"defective" biconditional table;*
3′. $(C||_{Fi}A)$*, the 2 × 2 Finettian interpretation of 3;*
4. $(C|_{?}A)$ *, the* 3 × 3 *general (underspecified) conditional;*
5. $(C|_{Fi})$ *, the* 3 × 3 *de Finetti conditional table;*
6. $(C|_{Fa}A)$ *, the* 3 × 3 *Farrell conditional table;*
7. $(C|_{C}A)$ *, the* 3 × 3 *Cooper conditional table.*

transitory state of *ignorance* (at a given time) until the statement is verified or falsified. Until this takes place, it is impossible to give it a truth value. Even though he did not vary in this conception of the third value, he used various terms to designate it; his favorite expression was "void" (e. g., de Finetti, 1967, 1974, 1995/2008) which we will adopt and will denote by "∅".

To formalize these notions, de Finetti (1936/1995, 1967, 1974, 1995/2008, 2006) defined a three-valued system that uses the third value *void* and is *superimposed* on a two-valued logic that uses *true* and *false*. We describe below the three-valued truth tables that define this system, specifying how these values are propagated for the usual connectives.

The second epistemic level in de Finetti (1980), *Level 2*, is a development of the first level. At this level, the initially non-numerical degrees of belief are finally expressed as numerical probability judgments. People are seldom fully ignorant about an event. They have expectations, make subjective probability judgments, engage in wagers, etc. This level corresponds to the full range of subjective degrees of belief about events where the initial ignorance and the ensuing uncertainty give way to the expression of additive probabilities. Fine distinctions are thus possible at Level 2, which is of psychological importance, since both ordinary people and scientists do often distinguish between events that are uncertain, judging some as more probable than others conditionally on their personal state of knowledge (Baratgin, 2015).

A substantial amount of research on uncertain reasoning has been carried out at Level 2—in fact most of the work mentioned above on the probability of conditionals or deduction under uncertainty. Hardly any research has been done to investigate Level 1 (with the exception of Baratgin et al., 2013, considered below). Level 1 should support and lead up to Level 2, and yet most contemporary theorists in the de Finetti tradition have concerned themselves with a much more refined and expressive system at Level 2 in which the third value for *if A then C* is specified by the conditional probability itself, *P(C|A)*, and the logical values *true* and *false* are replaced with 1 and 0 (Gilio, 1990; Jeffrey, 1991; Coletti and Scozzafava, 2002; Pfeifer and Kleiter, 2009). As Baratgin et al. (2013) point out, Level 2 removes some anomalies in Level 1, for people are never ignorant of trivial tautologies, such as *if A & C then A* and *A or not-A* (Over and Baratgin, 2017). But people do not always, and could not always, make such fine-grained evaluations of Level 2. They can, however, simply express their ignorance, or in other words, can remain at the transitory level 1. In summary, there is a gap to fill. De Finetti's theory has gained much experimental support at Level 2, but the question of its descriptive adequacy at Level 1 is open. The present paper addresses this question in several experiments, our aim being to test the descriptive adequacy, for ordinary people's judgments, of de Finetti's Level 1 in his overall theory of subjective probability. For half a century research in the psychology of reasoning has produced robust results on the comprehension of the connectives of propositional logic. People's performance indicates that they possess negation and conjunction, and to a lesser extent, disjunction (Manktelow, 2012) but their comprehension of the material conditional and biconditional is "defective," in the sense mentioned above. However, these studies were limited to the framework of classical bi-valued logic. The change of conceptual framework brought about by the Finettian theory necessitates that these studies be carried out with a tri-valued logic. The present study applies itself to refine and reinterpret the old results.

We now turn to the analysis of de Finetti's three-valued system in some detail. We begin with focusing on the conditional, which leads us to the concept of *conditional event* (or tri-event). A conditional event is defined by de Finetti (1936/1995) as a logical entity that is true when the antecedent A and the consequent C are true; false when A is true and C false; and void in the sense introduced above when A is false. The conditional event is closely analogous to a conditional bet, which is won in the first case, lost

in the second case, and called off in the third case, when it is "void" and no one wins or loses (see Politzer et al., 2010, on this analogy and the relation to Ramsey, 1926/1990, 1929/1990). So, at Level 1, the betting interpretation helps illustrate the void case[2].

It now appears that the empirical "defective" table for the conditional should be called the "2 × 2 de Finetti table" (and similarly for the "defective" biconditional) to avoid the negative term "defective" (Milne, 2012; Baratgin et al., 2013, 2014; Nakamura and Kawaguchi, 2016). Notice (**Table 1**) that in columns 1 and 3 the empirical "defective" table bears a symbol "I" (for *irrelevant*), whereas in column 1′ and 3′ the 2 × 2 de Finetti tables bear the symbol "∅" (for *void*). This point deserves explication. The 2 × 2 "defective" table in column 1 (**Table 1**) describes the psychological observation that participants judge that the states of affairs in which the antecedent is false do not allow to evaluate the conditional sentence in terms of true or false. Participants say that the sentence is neither true nor false, or that it may be true or false, or that one cannot know, and the term "irrelevant" (readily endorsed by participants) was coined by psychologists to express participants' perplexity about the truth value of the sentence. In other words, "irrelevant" and "void" refer to the same state of ignorance, the former being empirically-based and descriptive, and the latter theoretical.

The next step is to take into account the ignorance that can affect elementary events, considering that they, too, can be true, false or void (because for de Finetti all events are conditional), leading to a three-valued (3 × 3) truth table for the conditional event (denoted by $C|_{Fi}A$ in **Table 1**, column 5) which de Finetti (1936/1995) called "subordination". Similarly, he defined three-valued truth tables for the ordinary connectives (negation, conjunction, disjunction, see below). This set of truth tables which we will call "3 × 3 de Finetti tables" constitutes *de Finetti's Level 1 system,* abbreviated to Fi.

Traditional psychological experiments on the "defective" table were limited by the fact that the antecedent *A* and consequent *C* of the conditional did not have the third value, but de Finetti's 3 × 3 table, while encompassing the 2 × 2 de Finetti table, allows *A* and *C* to have the third value. In brief, we can find in the 3 × 3 de Finetti tables an answer to the question of what value does *if A then C* have when *A* or *C* have the third value (even though this was not his main objective). Of course, because the 3 × 3 table incorporates the 2 × 2 table, it keeps answering the question of what value does *if A then C* have when *A* is false: it is void. Note that *void* defined by a state of ignorance (as well as *irrelevant* expressed by participants in psychological experiments) is not a truth value homogeneous with *true* and *false*; rather, it is a meta-evaluation (for an analysis of this point, see Dubois and Prade, 2001; Dubois, 2008). It is in this sense that the 3 × 3 logic is superimposed on the 2 × 2 logic.

The conditional is so important that Baratgin et al. (2013) initially focused on it in their experimental study of three-valued tables. They observed that almost 60% of participants who gave responses in agreement with de Finetti's 2 × 2 table expanded it to produce de Finetti's full 3 × 3 conditional event table, when evaluating indicative conditionals and conditional bets. This is the first result supporting de Finetti's Level 1 system, but it is limited. Extending it to the other connectives of the system would demonstrate its descriptive adequacy, that is, *provide a semantic theory of the interpretation of natural language connectives under uncertainty.* This is the objective of the present paper. But before proceeding to the experiments, we should make some theoretical and methodological points. There exist many three-valued logical systems (for reviews, see Rescher, 1969; Haack, 1974; Gottwald, 2015). Some of them appeared before de Finetti, and many more have appeared in cognitive science since then. Psychologists of reasoning have so far done little to study whether any of these tables matches the judgments of ordinary people when they are in a state of ignorance about what is true and what is false (but see Elqayam, 2006, on "liar" paradoxes)[3]. Some of these systems propose a conditional table encompassing the 2 × 2 de Finetti table and so constitute possible alternative theories to de Finetti's Level 1 system, Fi. We give a short overview of these systems in the next section. See **Appendix A** (Supplementary Material) for a more detailed description, and **Appendix B** (Supplementary Material) for a presentation of the authors' individual reasons for developing their systems.

NINE SYSTEMS OF THREE-VALUED TABLES

An Extension of 2 × 2 Bi-valued Tables

In addition to de Finetti's (1936/1995) 3 × 3 table for the conditional event, there exist numerous other possibilities to build a 3 × 3 table to represent the indicative conditional of natural language, which we will call the *natural conditional.* Consider column 4 in **Table 1** in which $C|_?A$ represents a general 3 × 3 conditional table for this natural conditional. Here *A* and *C* can be true ("T"), false ("F"), or judged to be neither. After lines 1, 3, 7, and 9 have been filled in with the values of the "defective" table, there remain five cells marked with "?." The basic question is: what value should be in the place of each "?" to represent the natural conditional? There are 243 possible ways (3^5), in theory, of completing this conditional table. The same question is also posed for the other connectives. Among the existing three-valued logics we have found only nine three-valued systems that extend the 2 × 2 de Finetti table for the conditional and that also propose 3 × 3 tables that extend standard two-valued logic for the conjunction and disjunction connectives. By "extending," we mean 3 × 3 tables that have the same *true* or *false* values as their 2 × 2 counterpart in lines 1, 3, 7, and 9 mentioned above. Looking for such extensions is motivated by the experimental evidence that the classical 2 × 2 conjunction and disjunction truth tables are produced by a majority of people (Manktelow, 2012).

[2]There is much more in the betting scheme: de Finetti (1937/1964) proposed it at Level 2 as a procedure to operationally evaluate P(C|A), from which it follows that the conditional event can be represented as a three-valued random quantity taking on values 1, 0, P(C|A) (Gilio, 1990). Note that de Finetti (1962, 1964/1972, 1974) proposed also the penalty criterion (based on the Brier score) as a procedure to operationally evaluate P(C|A) (see for a recent analysis Gilio and Sanfilippo, 2011).

[3]Stenning and Van Lambalgen (2008) used Kleene's (1938) three-valued logic in the framework of logic programming, but they did not study participants' truth tables.

These nine systems of three-valued tables originate from the work of logicians, linguists, philosophers, and artificial intelligence researchers, who had different theoretical interests and approaches. As we will see in section Interpreting the Connectives and **Appendix B** (Supplementary Material), this is most evident in their interpretation of the third value. Some of these systems were not originally intended to represent an intuitive sense of uncertainty, which de Finetti aimed to capture (Baratgin and Politzer, 2016), but even so, they do have some *prima facie* interest for psychological modeling, simply because they extend the traditional 2 × 2 tables of two-valued logic to three-valued systems. Three-valued judgments have long been found in truth table studies of the conditional in psychological research, as we have described.

In summary, there are four basic connectives (negation, conditional, conjunction, disjunction). Three types of conditional (see **Table 1**, columns 5, 6, and 7) and four types of conjunction and disjunction (see **Appendix A**, **Table A.2** in Supplementary Material) constitute the *differential* building blocks of the nine three-valued systems: as displayed in **Appendix A**, **Table A.3** (Supplementary Material), each system is defined by using the involutive negation and by selecting one type of connective among the other three basic connectives[4]. A short reminder on the origins of three-valued logic is given in **Appendix B** (Supplementary material), followed by the origins of the nine "extended" systems [numbered (1)–(9)].

Interpreting the Connectives

The different truth tables for the connectives in **Table 1** and **Tables A.1**, **A.2**, **A.5** (Supplementary Material) may appear somewhat formal, and so we give a brief informal overview of how they differ from each other. We begin with the conditional *if A, then C*.

Recall that six systems, (1)–(6) in **Appendix B** (Supplementary Material), adopt the Fi conditional and so share the notion that a conditional with a false antecedent takes on the value ∅. Indeed, we have already seen through the betting schema that, whenever the antecedent A is not known to be true (∅ or F), the Fi conditional takes on the value ∅. In addition, a conditional sentence whose antecedent is true takes on the truth value of its consequent.

What distinguishes the Fi conditional from the other two conditionals, in columns 6 and 7 of **Table 1**, appears precisely for the value ∅ of the antecedent in lines 4 and 6. Two of the nine systems, (8) and (9) in **Appendix B** (Supplementary Material), use the Cooper conditional. For this conditional, with a truth-value gap ∅ (denoted by G for gap by Cooper) for the antecedent, the conditional takes on the value of the consequent, which is also the case when the antecedent is T. This captures the notion that the conditional has the same value with a ∅ antecedent as it has with a T antecedent. Only when the antecedent is F is the conditional ∅ whatever the value of the consequent. The Farrell conditional, (7) in **Appendix B** (Supplementary Material), differs in that it adopts a slightly more cautious evaluation: When the antecedent has a truth-value gap ∅ (denoted by I for ignorance by Farrell) and the consequent is T the conditional is not T but ∅ (**Table 1**, column 6). Note that both concur in holding the conditional to be F when the antecedent is ∅ and the consequent F. How this differs with Fi can be exemplified as follows. Suppose it is unknown whether *this chip is square*, while it is false that *this chip is black*. Then to evaluate *if this chip is square, then it is black*, some theorists (like Farrell and Cooper) may have the intuition that it is "false," whereas others (like de Finetti) may have the intuition that the value is "void"[5].

We can further examine the differences between systems by comparing the four types of conjunction and disjunction on which they are based that we have identified, viz., KLH, B, S, and M [defined in **Appendix A** (Supplementary Material)]. Most proposed systems (like Fi) in **Appendix B** (Supplementary Material) have truth-value gaps and consequently differ from three-valued systems proper in which the third value is homogeneous with the values T and F to which it can be compared using a relation of order. Most authors define an order between the truth-value gap and T and F. In de Finetti's framework, the truth-value gap is viewed as intermediate between F and T. Mura (in de Finetti, 1995/2008) gives a pragmatic justification with the bet schema: the payoff of a void bet is clearly intermediate between the payoff of a bet that is lost and a bet that is won (for more technical justification, see Hailperin, 1996; Milne, 1997, 2004; Blamey, 2001; Mura, 2016). It is exactly the order of KLH connectives. Conjunction obeys the following principle: the three values are formally put in an order denoted by $F \leq \varnothing \leq T$ (Dubois and Prade, 1994); then, whenever two sentences are connected, the value of the conjunction is the minimum of their values, that is, the conjunction gets the "weaker" value. Consider a context of chips of different shapes and colors. With the interpretation of ∅ as a truth-value gap resulting from ignorance, take a true sentence, for instance *the chip is square* (T), and suppose one is ignorant whether *the chip is black* (∅); then the conjunction *the chip is square and it is black* is evaluated as ∅ because min(T, ∅) = ∅. Suppose now *the chip is square* to be false; then the conjunction *the chip is square and it is black* is evaluated as F because min(F, ∅) = F. Similar considerations obtain for disjunction, *mutatis mutandis*. Here the value of the connection is defined by the maximum values of the disjuncts. If it is known to be true that *the chip is square* (T) and one is ignorant whether *the chip is black* (∅), then *the chip is square or it is black* will be evaluated as true because max(T, ∅) = T.

The various conjunctions obey the min order but they have their own formal order for the three values, which in

[4]With the exception of the R system, the material conditional and the material biconditional are not defining features because they can be derived compositionally from the basic connectives using the formulas $A \supset C =_{df} \neg A \vee C$, and $A \Longleftrightarrow C =_{df} (A \supset C) \wedge (C \supset A)$.

[5]One may balk at this notion because conditionals typically have uncertain antecedents and nevertheless they often convey a high degree of belief (or disbelief) rather than ignorance. One need not know whether it is true that *this man will fall from the 20th floor* to hold it to be false that *if this man falls from the 20th floor he will survive*. In this apparent counter example, which is on Level 2, the common knowledge suggests a degree of belief. In contrast, with our abstract and arbitrary material on Level 1 individuals have no expectations about the truth value of the conditional.

fact differentiates them from each other. The same obtains for the disjunctions with the max order. We will not review them in detail, but will have a look at what the choice of an order intuitively means. Take Bochvar's (1938/1981) conjunction $\wedge_B$. Its order is $\varnothing < F < T$. This means that whenever a sentence with the third value is connected (conjunctively) with another sentence whatever its value, the third value prevails and "contaminates" the conjunction. Similarly for disjunction $\vee_S$, the order corresponds to $F < T < \varnothing$. For instance, with a third value interpreted as *of no interest,* the sentence *the chip is square* being true or false and *the chip is black* being of no interest, the sentence *the chip is square or it is black* will be evaluated as being of no interest in each case. The situation is opposite for the Sobocinsky connectives where the orders are $F < T < \varnothing$ for conjunction and $\varnothing < F < T$ for disjunction, resulting in connections that appear to be "immune" to the third value as the other values absorb it. With the previous example, the disjunction will be evaluated as T in the first case and F in the second one.

Finally, consider involutive negation, which all the systems share. T is negated by F and F by T like in two-valued logic. Negating the third truth value by itself captures the intuition that one cannot consider a sentence that is not $\varnothing$ as T any more than consider it as F, so that it remains $\varnothing$.

Prima facie all the nine systems, irrespective of their origins and motivations, provide candidates for three-valued tables relevant to the psychological modeling of people's comprehension of connectives under uncertainty. They accommodate and extend the 2×2 de Finetti table for the conditional, which is supported by earlier psychological research, as we have explained. Most of the systems above are directly relevant to psychologists, especially those motivated by linguistic considerations and the inappropriateness of the material conditional to represent people's interpretation of the natural language conditional (such as BFM, etc., defined in **Appendix A** (Supplementary Material). Clearly, an empirical investigation is necessary to decide which of these three-valued systems best fits ordinary people's judgments about natural language connectives. We present several experiments that aim to answer this question by examining people's truth tables for negation, conjunction, disjunction, the conditional, and the biconditional. More strongly, we ask whether the tables closest to people's judgments belong to one system in the literature. For all the reasons detailed in section The de Finettian Approach, and in view of the results we have already obtained for the conditional, we consider de Finetti's Level 1 system as the most serious contender. Recall that it is characterized by the Fi conditional and the KLH conjunction and disjunction.

EXPERIMENTS: THE FINETTIAN AND OTHER THREE-VALUED SYSTEMS

Method

Participants

In Experiment 1 ($N = 54$) and Experiment 2 ($N = 101$), participants were French native speakers. They were students at the University Paris 8 who volunteered for the experiments. They already held a degree and were resuming their studies in a remote teaching program in the social sciences. They had no specific background in logic or probability theory. In Experiment 3, participants were 58 undergraduate Japanese native speaker students enrolled in a computer programming class at the Tokyo Denki University. All were naive to the purposes of the study. Experiments 1 and 2 were administered on a computer screen and Experiment 3 was presented in a booklet. An online informed consent was obtained from all participants. This study was carried out in accordance with the recommendations of the APA ethical principles and code of conduct and was approved by the ethics committee of Laboratoire Cognitions Humaine et Artificielle (*EA 4004–CHArt*), Université Paris 8, France.

Materials

In Experiment 1 and 2 the same material was used. Participants were presented with sentences that referred to a chip that could be in one of two colors, black or white, and one of two shapes, square or round. The task was to judge whether the sentences were true, false, or neither. There were different conditions of visibility. In one condition, the chip was seen through a transparent window, making it clearly true whether the chip was square, or round, and similarly making it clearly true whether the chip was black or white. In another condition of visibility, the chip was seen through a device that made it visually impossible to know whether the chip was square or round. And in a third condition of visibility, the chip was seen through a filter making it visually impossible to know whether the chip was black or white. And finally the chip could be seen through both the device and the filter, making both the shape and the color impossible to identify. This technique allowed us to fill up the nine cells of a three-valued truth table with the participants' responses (see **Figure 1**).

In the third experiment (Japanese participants), an isomorphic material with pictures of round or pointed chips that could be blue or red was used (**Figure 2**).

Design and Procedure

In the three experiments, participants were required to judge the truth value of the sentence under consideration for the nine combinations corresponding to the nine cells of the truth table (see **Figure 3**).

For the three experiments the combinations were presented in a random order and each one was accompanied by three response options: *certainly true, certainly false, neither true nor false.* The participants were required to select one option (see an example in **Figure 4** for the conjunction).

The choice of the adverb "certainly" reflects the Finettian notion that when an event is known to have occurred or not to have occurred, this is known with certainty, and so the truth or falsity of the proposition that expresses it is certain. Besides, this should avoid possible common fuzzy interpretations of "true" and "false" such as "very likely to be true/false." This wording has already been used for the same purpose in the context of research on the framing effect (Mandel, 2014).

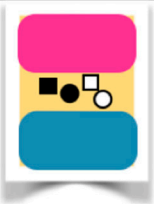
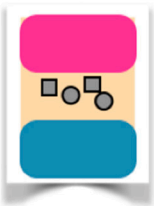
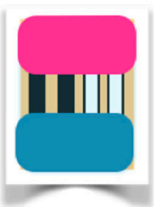
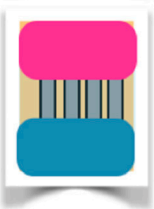
Statement	Photographs
A pink opaque box is filled with chips that can be in one of two colors (black or white) and in one of two shapes (round or square). Chips chosen at random are dropped into an opaque blue box situated underneath. Whenever one or several chips are dropped into the blue box a photograph is taken instantaneously. For example, the photograph on the right has been taken after four chips have been chosen (one chip in each shape and each color).	
-The photograph can also be taken using a color filter that makes the chips look grey, whether they are black or white, showing only their shape. The photograph on the right was taken after four chips had been dropped (one chip in each shape and each color).	
- The photograph can also be taken using a device that conceals the shape of the chips, whether they are round or square, showing only their color. The photograph on the right was taken after four chips had been dropped (one chip in each shape and each color).	
- Finally, the photograph can also be taken using both the color filter and the device that conceals the shape. In that case the falling chips leave a grey track. The photograph on the right has been taken after four chips had been dropped (one chip in each shape and each color).	

FIGURE 1 | The presentation of the game in Experiments 1 and 2.

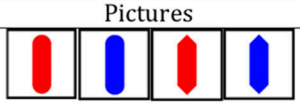
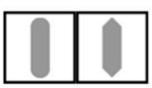
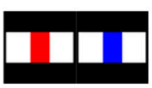
Statement	Pictures
The pictures represent a chip that is round or pointed in shape and red or blue in color, yielding four variants in total.	
A picture can be developed in black and white, concealing the color of the chip. (There are no grey chips.)	
A picture can be damaged on the upper and lower ends, concealing the shape of the chip.	
A picture can be damaged and monochrome. Then both the shape and color of the chip are uncertain.	

FIGURE 2 | The presentation of the game in Experiment 3.

The choice of the wording *neither true nor false* for the third option was made for several reasons. First it should be as close as possible to de Finetti's conception and formulation of the third value. This third value (or *void,* as de Finetti often called it) is the evaluation made by an individual who is not in a position to know whether an event is true or false. Commenting on the tri-event, de Finetti explicitly states that the third value is to be regarded as neither true nor false: "Whenever the *condition B* is satisfied, then $A|B$ is either true or false (1 or 0). But unless the condition B is satisfied, one can neither say that the event $A|B$ is true, nor that

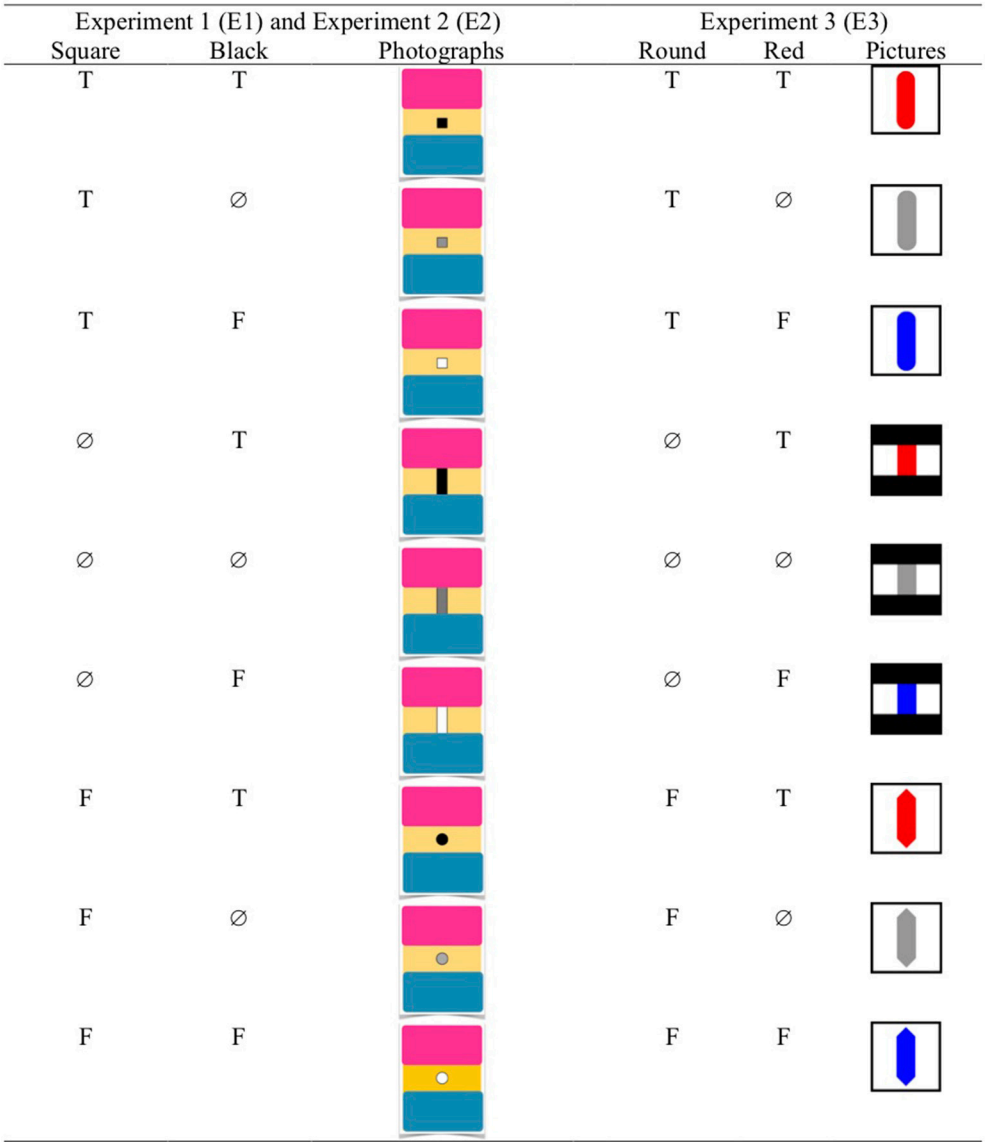


FIGURE 3 | The nine possible combinations of photographs (Experiments 1 and 2) or pictures (Experiment 3) corresponding to the nine cells of the three-valued truth table.

the event $A|B$ is false. It is *void* or *null* in the sense that the premise under which it is considered either true or false no longer holds. In my opinion, these three cases should be treated as distinct" (de Finetti, 1995/2008, p. 170).

Second, it should capture participants' natural evaluations, that is, with as little suggestion as possible. In principle, the third option could be "true or false" (or some equivalent expression such as "It could be true or it could be false." This option is correct (and trivial) from a logician's objective point of view. But it does not readily accommodate subjective judgments, in particular those generated by three-valued systems (see the various and subtly different interpretations of the third value in section Interpreting the Connectives). By parity of argument, it might be objected that "neither true nor false" cannot accommodate the choice of "true or false" because it is incompatible with it. This is correct, but pragmatically rejecting the assertions that the sentence is true and that it is false gives rise to the assertion that it is neither. More precisely, participants who do not find an assertable option are led to interpret the third option as a means to express just this (and to disregard the logical triviality in case it had come to their mind). The judgment that neither "true" nor "false" are adequate options induces the judgment that "neither true nor false" is adequate, which turns the third option into a meta-option equivalent to "other" that cannot be put on the same level as "true," "false," and "true or false."

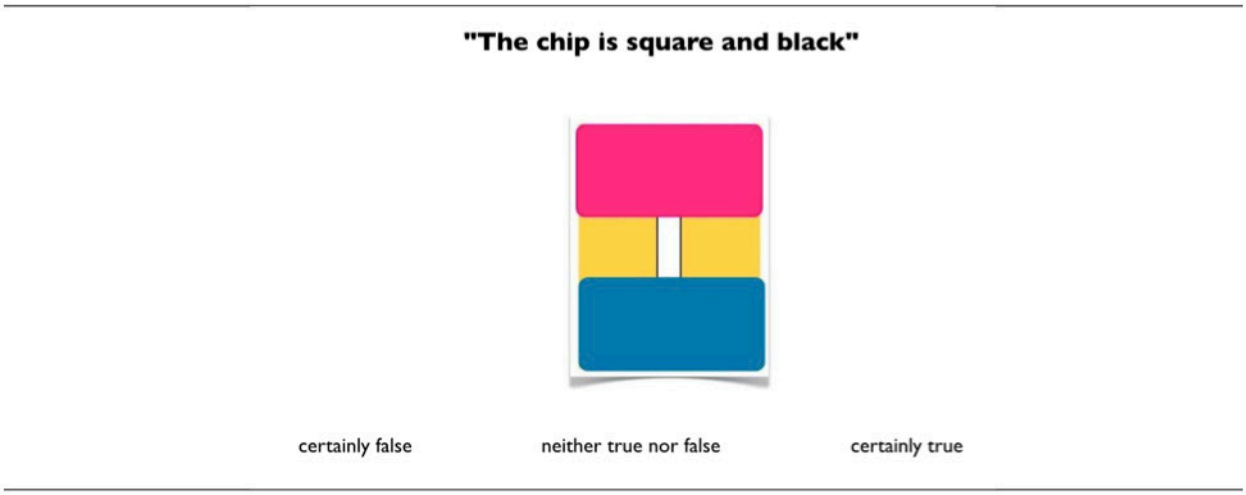


FIGURE 4 | Example of a trial (one logical combination) for conjunction. The first conjunct has the third value ∅ and the second conjunct is false.

Third, the format should be common to all the connectives. For the conditional in particular, it should be possible for participants to express a judgment such as "void" or "irrelevant" without any suggestion, which the "neither" option satisfies. Note that the first constraint above is exemplified with the conditional which theoretically returns the value *void* in case its antecedent is not known to be true, but the other connectives also have logical cases of voidness for which the option "neither" is appropriate for the same reasons. In brief, the aim of the third option is to capture the judgment that neither the first option nor the second is adequate, in the spirit of de Finetti, without influencing the participants, while being applicable to the various connectives, and the formulation adopted does just that.

In Experiment 1, each participant was asked to judge the truth value of a negated sentence (e.g., *the chip is not square* when the shape of the chip presented could be square or round or indeterminate, and the color black, white or indeterminate), hence nine presentations (or "trials"). Participants were randomly allocated to one of the two statements, *the chip is not square* and *the chip is not black*.

In Experiment 2, each participant received four sentences: first the simple affirmation, *The chip is a square* to familiarize them with the task. This was followed by a conjunction, *The chip is square and black*; then there were two sentences presented in a counter-balanced order: a disjunction disambiguated by "or both" written in parentheses, *The chip is square or black (or both)*[6], and a conditional, *If the chip is square, then it is black*. The conditional will not be detailed here (for the results, see Baratgin et al., 2013).

In Experiment 3 (Japanese sample), each participant received four sentences, in this order: the simple affirmation, *The chip is red*, the conjunction, *The chip is round and red*, the conditional, *If the chip is round, then it is red*, and the biconditional, *If the chip is round, then it is red, and if it is red, then it is round*. The original sentences in French and Japanese can be found in **Appendix C** (Supplementary Material).

Results

Method of Analysis

The tables produced by participants will be analyzed in two stages. The first stage casts the results in terms of the traditional two-valued classification. That is, we restrict the analysis of the answers to the four "old" cells of the traditional table that correspond to the four cases where the antecedent and the consequent are either true or false. This allows the identification of a 4-cell truth table for each connective and each participant (and a 2-cell truth table in the case of negation). In this way, we take up the classic 2×2 tables before extending them into new 3×3 tables.

In the second stage of the analysis, we further characterize the tables by considering all nine cells (and all three cells for negation). Then the observed three-valued tables are compared with the relevant three-valued formal tables of the nine systems.

The First Stage Analysis

Table 2 displays for each connective the frequency distribution of the interpretations (the tables produced) in percent. To answer the research question, we were basically interested in the identification of the modal response, that is, we were looking for a dominant interpretation belonging to the same system across connectives. In each of the first three columns there is one modal response >70% (close or equal to 100% in the first three columns), that is, a clearly dominant response appears. However, in the last two columns (conditional and biconditional) the modal response is not so high. To identify this modal response as a reliable dominant interpretation, a 95% confidence interval for proportions (based on z values) was calculated (rounded to the closest unit) for all percentages >10%. Confidence intervals will also be given for the second stage analysis.

[6]In doing so, we followed psychologist's traditional way of disambiguating "or" in the study of reasoning.

TABLE 2 | First stage analysis.

Connective Tables produced	Negation E1	Conjunction E2 and E3	Disjunction E2	Conditional E3	Biconditional E3
Negation $\neg A$ (*or* $\neg C$)	100 [94; 100]				
Conjunction A∧C		98 [93; 100] (Experiment 2) 98 [90; 100] (Experiment 3)		22.4 [14; 35]	20.7 [12; 33]
Disjunction A v C			73.3 [64; 81]		
Conditional "defective" $C\|_{"d"}A$				37.9 [27; 51]	1.7
Material conditional $A \supset C$				3.4	
Material biconditional $A \Leftrightarrow C$				15.5 [8; 27]	25.9 [16; 39]
Biconditional $C\|\|A$				15.5 [8; 27]	50.0 [38; 63]
Other		2 (Experiment 2) 2 (Experiment 3)	26.7 [19; 36]	5.1	1.7

Frequency distribution of the tables produced (in percent) for the five connectives considering only two truth values for A and C. In brackets: 95% confidence intervals. E1, Experiment 1, N = 54; E2, Experiment 2, N = 101; E3, Experiment 3, N = 58.

For negation (experiment 1), all participants answered in agreement with the two-valued truth table of negation.

For the conjunctive statement, 98% of the participants in Experiment 2 as well as in Experiment 3 respected the conjunction table.

For the disjunctive statement 73.3% respected the disjunction table. These rates correspond to the traditional rate of response presented in the literature. In particular, the review made by Evans et al. (1993) for disjunction shows that the true-false combinations are evaluated as false between 10 and 28% of the time, indicating a conjunctive interpretation. Similarly, virtually all of the 27 participants who did not respect the standard truth table answered *false* to the true-false combinations, either on one occasion (20) or on both (6). This means that these participants had difficulty processing disjunction and had a tendency to construe it as a conjunction in line with the classic results, and that their error was not due to having trouble with the uncertain cases or with the response format, that is, with the three-valued system. Finally, there was no case of exclusive interpretation, indicating that the disambiguation by "or both" was effective.

For the conditional statement, the two main tables produced by Japanese participants of Experiment 3 correspond to the usual "defective" conditional (37.9%) and conjunction tables (22.4%). These frequencies are comparable to Baratgin et al. (2013) French data. The only notable difference is that the frequency of the biconditional table which was virtually null now reaches 15.5%.

For the biconditional statement, the dominant interpretation is the 2 × 2 de Finetti table (50%), followed by the material biconditional table (25.9%) and the conjunction table (20.7%).

The Second Stage Analysis

We consider all nine cells of the observed truth tables. Each participant's table is classified by considering the formal table to which it is the *closest*. Our criterion of "closeness" or "distance" is as follows. A participant's table is taken to be a perfect instance of a formal table X when it is identical to X. A participant's table is a "close" instance of X when it differs from X just by one cell, and *from any other formal table by more than* one *cell*. If a participant's table differs equally (by one cell) from two (or more) formal tables, it is still "close" to, but classified as *ambiguous* between, these tables (these are equally likely). Finally, if a participant's table differs by two or more cells from all formal tables, then it is classified as "indeterminate": it differs too much to make a reliable identification.

First of all, for the simple affirmation, *the chip is square,* all participants answered correctly, that is, *certainly true* when the chip was square, *certainly false* when it was round, and *neither true nor false* when its shape was blurred. This is evidence that the square, blurred, and round shapes were visually well distinguished, allowing participants to recognize the three logical possibilities, and in particular, the representation of uncertainty by the blurred image. Importantly, there was a perfect one-to-one correspondence between the blurred image and the *neither* answer, which validates this formulation.

For the negation, 48 participants (89%) fully conformed to the involutive negation $\neg_i$ (in which the third value maps onto itself) on all nine trials, and six participants (11%) answered in agreement with this table on eight trials (meaning that they were closer to the involutive negation than to any other type of negation). Two of these six participants clearly made a well-known slip triggered by double negation (Wason, 1959), answering F instead of T to a round chip when the sentence was *The chip is not square*. The other four participants negated the ∅ chip by answering F (three cases) or the T chip by answering ∅ (just for one case). The answers provided by these four participants are thus closest to an involutive negation than to left and right negations. In brief, we find evidence of only the involutive negation.

Before considering conjunction and disjunction, note that the numbers for these two connectives are smaller than they are in the first stage. This is because the three-valued tables for conjunction and disjunction (like for negation) are built as expansions of the corresponding classic two-valued tables which serve as filters, so that only participants who have produced

the latter can be considered in the second stage. For instance, we have mentioned earlier that 98% of the 101 participants in Experiment 2 (i.e., 99 participants) produced a conjunction table in the first stage analysis. Consequently, the second stage analysis for conjunction and the related percentages are based on those 99 participants. The only case where N is notably diminished is disjunction in Experiment 2 (from 101 to 74, as mentioned in section The First Stage Analysis). The results for conjunction and disjunction are detailed in **Tables 3**, **4** in which we will examine the sum column.

For conjunction (**Table 3**), it is apparent that a large majority of the observations coincide with the *KLH* connective $\wedge_K$ defined in **Appendix A, Table A.2** (Supplementary Material) (82.8% in Experiment 2 and 96% in Experiment 3). The remaining interpretations correspond to the McCarthy $\wedge_M$ conjunction (13.1 and 4%, respectively).

For disjunction (**Table 4**), the absolute majority of the observations (58.1%) coincide with the *KLH* connective $\vee_K$ defined in **Appendix A, Table A.2** (Supplementary Material). The remaining interpretations coincide with the Sobocinsky disjunction $\vee_S$ (14.9%) and to tables that we call "ambiguous" because they differ equally (by one cell) from both disjunctions $\vee_K$ and $\vee_S$.

For the conditional (**Table 5**), almost all the participants' interpretations coincide with a table that belongs to the Fi system. We find notably that all of the 22 participants whose first stage table was identified as de Finetti's 2×2 "defective" table expanded this table into de Finetti's conditional event (3×3) table. In other words, the conditional probability to produce the three-valued conditional event table knowing that the two-valued table is the "defective" table equals one. Also of interest is the fact that most participants (84.6%) who have a conjunctive interpretation of the conditional in the first stage expand this table into a conjunction table (the KLH table) that is in the Fi system. These results confirm the observations of Baratgin et al. (2013) with French participants. Similarly, most participants giving a biconditional interpretation produce the Fi biconditional. Interestingly, even for the material biconditional interpretation, most participants produce the associated Finettian table.

For the biconditional sentence (**Table 5**), the observations are identical: almost all the participants' interpretations coincide with a table that belongs to the Fi system. In particular, the dominant biconditional interpretation is always the Finettian one, that is, 100% of 3×3 de Finetti biconditional table. Similarly, most participants (86.7%) with a material biconditional interpretation choose the expanded Kleene material biconditional [defined in **Appendix A, Table A.5** (Supplementary Material)] and also most of those (83.3%) with a conjunctive interpretation produce the associated Finettian table [the KLH conjunction $\wedge_K$ defined in **Appendix A**, **Table A2** (Supplementary Material)]. All this suggests a remarkable consistency within a unique logical system, namely the Fi system.

We can summarize these results as follows. The overwhelmingly dominant table for *A and C* is the KLH conjunction $\wedge_K$ and the dominant table for *A or C* is the KLH disjunction $\vee_K$, both of which are features of the Finettian system. Whatever the interpretation for *if A then C* (conditional, conjunction, biconditional, material biconditional), it is the corresponding Finettian table that is overwhelmingly the dominant choice. This obtains also for *if A then C and if C then A,* whatever its interpretation (conjunction, biconditional, material biconditional). In addition, the involutive negation is always observed.

DISCUSSION

De Finetti's Level 1 System as the Best Approximation

The hypothesis that de Finetti's Level 1 system is adequate to model the psychological three-valued truth tables for natural language connectives is clearly supported by the results in the following two respects. One, its constitutive connectives: involutive negation $\neg_i$, the KLH conjunction $\wedge_K$, the KLH disjunction $\vee_K$, the Fi conditional $C|_{Fi}A$ and the Fi biconditional $C||_{Fi}A$, have been found to be the dominant interpretations.

TABLE 4 | Second stage analysis. Disjunction.

Disjunction tables produced	(0)*	(1)**	Sum (0)+(1)
KLH ($C\vee_K A$)	52.7	5.4	58.1 [48; 69]
Sobocinsky ($C\vee_S A$)	8.1	6.8	14.9 [9; 25]
Ambiguous (1 difference with $\vee_K$ and with $\vee_S$)		16.2	16.2 [12; 29]
Other			10.8 [8; 22]

*Frequency of tables produced (in percent) considering three truth values. Experiment 2, N = 74. *(0), 0 difference (all nine cells coincide); **(1), one difference (8 cells coincide). In brackets: 95% confidence intervals.*

TABLE 3 | Second stage analysis. Conjunction.

Conjunction tables produced	(0)*		(1)**		Sum (0)+(1)	
	E2	E3	E2	E3	E2	E3
KLH ($C\wedge_K A$)	76.7	96	6.1	0	82.8 [74; 89]	96 [88; 99]
McCarthy ($C\wedge_M A$)	13.1	4	0	0	13.1 [8; 21]	4
Other	4.1				4.1	0

*Frequency of tables produced (in percent) considering three truth values. *(0), 0 difference (all nine cells coincide); **(1), one difference (8 cells coincide). In brackets: 95% confidence intervals. E2, Experiment 2, N = 99; E3, Experiment 3, N = 57. The Table reads as follows: in Experiment 2, 76.7% of the 99 participants produced the exact KLH table, and 6.1% produced it with one difference, so that 82.8% produced the KLH table with at most one difference, with a 95% confidence interval of [74; 89], etc.*

TABLE 5 | Second stage analysis. Conditional and biconditional.

	Conditional			Biconditional		
Tables produced	**(0)***	**(1)****	**Sum (0)+(1)**	**(0)**	**(1)**	**Sum (0)+(1)**
Conditional			*N = 22* (37.9%)			*N = 1* (1.7%)
de Finetti ($C|_{Fi}A$)	95.5	4.5	100 [85; 100]	100		100
Conjunction			*N = 13* (22.4%)			*N = 12* (20.7%)
KLH ($A \wedge_K C$)	53.8	30.8	84.6 [58; 96]	75	8.3	83.3 [55; 95]
Other	15.4		15.4 [4; 42]	16.7		16.7 [5; 45]
Material conditional			*N = 2* (3.5%)			
Kleene ($A \supset_K C$)	50		50			
Other	50		50			
Material biconditional			*N = 9* (15.5%)			*N = 15* (25.9%)
Kleene ($A \Leftrightarrow_K C$)	77.8	11.1	88.9 [56; 98]	66.7	20	86.7 [62; 96]
Other			11.1 [2; 44]	13.3		13.3 [4; 38]
Biconditional			*N* = 9 (15.5%)			*N = 29* (50%)
de Finetti ($C||_{Fi}A$)	77.8		77.8 [45; 94]	93.1	6.9	100 [88; 100]
Other			22.2 [6, 55]			
Other			*N* = 3 (5.2%)			*N = 1* (1.7%)

*Frequency of tables produced (in percent) considering three truth values. Experiment 3, N = 58. *(0), 0 difference (all nine cells coincide); **(1), one difference (8 cells coincide). In brackets: 95% confidence intervals. The Table reads as follows: for the conditional, 22 participants (out of 58 = 37.9%) produced a conditional table that was identical to de Finetti's table and no other conditional table was observed; still for the conditional, 13 participants (out of 58 = 22.4%) produced a conjunction table; 11 of these (84.6%) produced a KLH table; and 2 (15.4%) produced a different conjunction table, etc.*

This was the case for the two languages studied, French and Japanese, which offers a remarkable cross-linguistic support to the Finettian theory on Level 1, given the remoteness of the two linguistic families. Two, even when the conditional and biconditional sentences are not construed as a conditional or a biconditional, respectively, the truth table that is produced still belongs most generally to the Fi system. However, it can be objected to the first point that the other two logical systems that are built on the same connectives, namely McDermott, and Reichenbach could, *eo ipso*, be regarded as possible candidates. Is there a way to decide between the three systems? We have seen earlier that the latter two differ from de Finetti in that they have additional connectives.

Consider first Reichenbach's system, (2) in **Appendix B** (Supplementary Material). It has additional connectives (two more negations), and two material conditionals and two material biconditionals [see **Tables A.1 and A.4** (Supplementary Material)]. We made no observation of a form of negation other than the involutive one, nor did we find any trace of the two forms of material conditional or biconditional. We can conclude that Reichenbach's three-valued logic is inadequate in that it predicts several truth tables, that is, interpretations of the negation, conditional, and biconditional, that our participants never had. This is not too surprising given that the objective of his logic is to account for a problem that belongs to the epistemology of quantum mechanics. Even though there is striking overlap between his system and the three-valued table of the Finettian conditional, the additional connectives needed for his purpose are irrelevant for psychological modeling. To take but one example of the lack of plausibility of the system from a psycholinguistic point of view, the cyclical negation of A requires a triple application of the operator to get back to A: $A = \sim\sim\sim A$; and the complete negation holds only as: $\overline{A} = \overline{\overline{\overline{A}}}$, whereas double negation does apply to diametrical (involutive) negation: $A = \neg_i\neg_i A$ (see Table A.1).

McDermott's system, (4) in **Appendix B** (Supplementary Material), also has additional connectives: one conjunction ($\wedge_S$) and one disjunction ($\vee_S$). For disjunction, we did find some trace of $\vee_S$ (15%, against 58% for $\vee_K$), but for conjunction we did not find any trace of $\wedge_S$. This does not support the system. However, before eliminating it, we must envisage that there may be special conditions or circumstances under which the second set of connectives is used, which our material may have failed to meet. McDermott (1996) contented himself to remark, based on intuition, that *and*, and *or* are ambiguous in natural language, hence his definition of two different connectives in each case. But to exemplify the ambiguity he did not use simple sentences made of two atomic components, such as *A and B,* or *A or B.* Instead, he used complex sentences, one component of which was always a conditional (such as *A and if B then C*, or *A or if B then C)*. Obviously, if this is required for the supplementary connectives to apply, the double connective claim cannot be refuted by our experimental results, which are based on at most two atomic sentences. McDermott's theory is not specified enough in its current state and the question remains open for further research. But it should be noted that if the claim becomes experimentally supported, it would come as an extension of the Finettian system proper. It is remarkable that McDermott's approach has much in common with de Finetti's, in particular in the assessment of truth values using the betting method, and crucially in the definition of the natural conditional. Finally, the conditions that trigger the additional connectives

interpretation could have a pragmatic explanation, keeping the Finettian system semantically unaltered. For all these reasons, the objections to the first point above seem hard to maintain; in addition they leave the second point unaffected. This is why we can confidently conclude that our results designate the Finettian system as the best approximation to the participants' three-valued truth tables obtained from judgments of truth and falsity of atomic sentences describing uncertain characteristics.

The Significance of the Results: Logic and the Study of Human Reasoning

Our results constitute a step toward giving an integrated answer to three related questions. One, is there a dominant interpretation of the basic connectives with sentences that have a truth-value gap? Two, do these interpretations constitute a consistent system? Three, is there a way to solve the half-a-century-old problem of the "defective" truth table of the conditional?

The Existence of a Dominant Interpretation

We have obtained an affirmative answer to the first question. For each connective (negation, conjunction, disjunction, conditional, and biconditional), participants' interpretations were distributed over a limited number of table varieties among numerous possible tables, and for each connective, there was a clear dominant interpretation, namely, involutive negation, the KLH conjunction $\wedge_K$, the KLH disjunction $\vee_K$, the Fi conditional $C|_{Fi}A$ and the Fi biconditional $C||_{Fi}A$, respectively. For negation there was a single interpretation (involutive). For conjunction the modal interpretation ($\wedge_K$), collapsed over two experiments, was close to 90%. For disjunction the modal interpretation ($\vee_K$) was chosen 58% of the time (among participants who had a 2×2 disjunctive interpretation) while the next most frequent interpretation was seldom chosen (16%). For the conditional and the biconditional, one interpretation ($C|_{Fi}A$ and $C||_{Fi}A$, respectively) was chosen 100% of the time (among participants who had the corresponding 2×2 interpretation). In brief, given people's two-valued interpretation, there is always one way to extend this interpretation to a three-valued table that musters an absolute majority, and (except for disjunction) there is near unanimity for this interpretation.

The Existence of a System

We have obtained an affirmative answer to the second question too: not only is there a predominant interpretation for each connective, but this interpretation always belongs to the same system. It could have been the case that while the dominant table for one connective belongs to one system, the dominant table for another connective belongs to another system. But this is not what we have observed: for each connective, out of all the possible tables, it is the one that belongs to the Finettian system that dominates. And there is more: even when the two-valued table has deviating interpretations, which occurs for the conditional and the biconditional, the table is almost always completed into the corresponding Finettian table. See for instance how in Experiment 3 the two-valued conjunctive interpretation of the biconditional made by 12 participants (20.7%, **Table 2**, first stage analysis) leads ten of them (83.3%) to the corresponding conjunctive Finettian ($\wedge_K$) three-valued interpretation shown in **Table 5** (second stage analysis). All this means that the present results are more than an extension to the other connectives of the results obtained for the conditional by Baratgin et al. (2013). Rather, what we have established here is *the existence, in people's judgments under uncertainty, of mutually consistent interpretations of the standard connectives organized in* one *system, namely de Finetti's Level 1 system*.

Interpreting the "Defective" Table

Finally, we have obtained confirmation of a positive answer to the third question, "Can the puzzle of the defective table be solved?" Three-valued truth tables generalize two-valued tables. They collapse into two-valued tables when the component sentences are certain. In such a case, for the conditional, the third value $\varnothing$ left in the body of the 2×2 table constitutes the "defective" table and the explanation of its origin. Note that there is no conflict between the two-valued and three-valued tables. The latter incorporate the former in the same way that rational numbers include integers.

The Significance of the Results: Interpreting the Third Value

We recalled some important findings in the introduction. For several decades psychologists have known that people judge that *if A then C* is true when *A* holds and *C* holds, false when *A* holds and *C* does not, and neither true nor false when *A* does not hold. For the last decade, there has been growing psychological evidence that people tend to judge that the probability of the indicative conditional, *P(if A then C)*, is the conditional probability of *C* given *A*, *P(C|A)*. There is also evidence supporting the claim that people tend to be coherent in explicit deduction under uncertainty. More recently, psychologists have shown that there is a close relation between indicative conditionals and conditional bets (Oberauer and Wilhelm, 2003; Politzer et al., 2010; Baratgin et al., 2013, 2014; Nakamura et al., 2018). There is an urgent need to integrate these experimental findings. The integration has been held back because psychologists did not raise the general question of which three-valued tables correspond most closely to people's judgments under uncertainty.

In the present paper, we have raised the question and proposed an answer based on de Finetti's Level 1 system, offering a model of the interpretation of natural language connectives under uncertainty. Obviously, we should keep in mind the limitations of our study due to the size of the samples and more importantly to the fact that the sentences in the experiments referred to specific materials. No overall investigation of the foundations of de Finetti's system, at Level 1, had yet been carried out. In view of the psychological relevance and plausibility of de Finetti's subjective approach to probability, and of the successful application of his concepts and ideas recalled above, it would have been deeply puzzling if the interpretation of connectives had been found to be at variance with his system. But on the contrary, our results showing that people conform to de Finetti's Level 1 system add much support to the project of developing the psychology of reasoning on a de Finettian basis, within the Bayesian account of ordinary reasoning that we are pursuing.

Our results should lead to more research. Important questions concerning the interpretation of the third value still await investigation. We have seen that the various systems of three-valued logic have different uses and objectives in defining a third value (even though most systems studied interpret it as a truth-value gap, see **Appendix B** (Supplementary Material). However, there appears to be an underlying common notion, that of doubt about truth and falsity due to uncertainty.

Even though we have identified the most frequent interpretation of each connective, namely the one that belongs to the Finettian system, it must be kept in mind that this result relies on a few experiments that operationalized uncertainty as ignorance about which of two values a visually defined variable had. Visual uncertainty about the identification of two shapes or two colors was hypothesized to coincide with de Finetti's concept of a "void" judgment. We certainly acknowledge that there is a need for additional experiments using other languages and, more importantly, that vary the source of uncertainty. There are perhaps other types of visual uncertainty, e.g., arising from soritical series (see Douven et al., 2018), and we should move beyond the visual modality, e.g., to sound or haptic modalities, and then beyond the sensory modalities, e.g., to logical or semantic uncertainty (as can be found in the paradoxes of self-reference; see Elqayam, 2006). The choices are unlimited, for uncertainty is everywhere in natural language, a point de Finetti himself would have emphasized. If the results of such experiments are consistent with the present observations, then there will be stronger support for the general conclusion that the Fi system is the best semantic theory of the interpretation of natural language connectives under uncertainty.

In contrast, it is possible to operationalize a much more common concept of uncertainty. Considering that ignorance reflects a lack of information that could be dispelled as information increases, still using the same material participants could be provided with frequency distributions about the proportions of round, square, black, and white chips, that is, manipulating the base rates. With this additional knowledge, one is invited to move from radical uncertainty to a gradable notion of uncertainty in which the individuals' degrees of belief vary between 0 and 1. In this modified situation there is no more total ignorance and the individual shifts from Level 1 to Level 2. But in doing so, one would be losing the state of total ignorance whose investigation is the objective of the present work, and as noted earlier (section The de Finettian Approach) the psychological research on reasoning under uncertainty (and indeed a large amount of research on judgment and decision making under uncertainty) has essentially been carried out on Level 2.

We have considered the basic connectives, but it might be interesting for future research to study more complex sentences than the basic ones. Conditionals can be embedded like in left embedding *If they were outside (O), then if it rained (R) they got wet (W)* or in right embedding *If the cup broke (B) if dropped (D), then it was fragile (F)* (see Gibbard, 1981; Douven and Verbrugge, 2013; Douven, 2016). In the Finettian framework of Level 1, these sentences can be written as $(W|_{Fi}R)|_{Fi}O$ and $F|_{Fi}(B|_{Fi}D)$, respectively and they collapse into a single form, $W|_{Fi}(R \wedge_K O)$ and $F|_{Fi}(B \wedge_K D)$, respectively (de Finetti, 1974, p. 328). Their truth tables can be established, allowing a further test of the theory on its Level 1 (van Wijnbergen-Huitink et al., 2015). Recently the Finettian treatment of embedded conditionals on level 2 has attracted the attention of theorists (Gilio and Sanfilippo, 2014; Douven, 2017; Sanfilippo et al., 2018) with results that reflect the difference of perspective between the two levels.

One final remark also for future research: in the current study we have compared various systems by eliciting judgments of truth value for connected sentences. Given that each system has a consequence relation, another way to test the systems against each other could be to study the elementary inferences that reasoners are willing to make.

AUTHOR CONTRIBUTIONS

JB and GP: design of the study and data analysis. JB and TT: data collection. GP and JB: draft of the manuscript. JB, GP, and DO: conceptual elaboration. DO and TT: critical revision of the manuscript.

FUNDING

This work was supported by the French ANR agency under grant ANR Chorus 2011 (project BTAFDOC) and by JSPS KAKENHI Grant Number 17H04696.

ACKNOWLEDGMENTS

We would like to thank Ikuko Hattori, Masasi Hattori, Yoshimasa Majima, Hiroko Nakamura, Jean-Louis Stilgenbauer, Hiroshi Yama, and Arai Yoshiko for much discussion and other help in our research and several reviewers of a previous version for helpful suggestions. We are also grateful to Junki Yokokawa for his help in running Experiment 3 and its analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01479/full#supplementary-material

REFERENCES

Baioletti, M., and Capotorti, A. (1994). "Two different characterizations of logical dependence on conditional events," in *Proceedings of the 3rd Workshop on Uncertainty Processing in Expert Systems (WUPES'94),* ed R. Jiroušek (Trešt), 7–17.

Baioletti, M., and Capotorti, A. (1996). "A comparison between classical logic and three-valued logic for conditional events," in *Proceedings of* Sixth *International Conference Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'96)* (Granada), 1217–1222.

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

Baratgin, J., Douven, I., Evans, J., S. B. T., Oaksford, M., Over, D., et al. (2015). The new paradigm and mental models. *Trends Cog. Sci.* 19, 547–548 doi: 10.1016/j.tics.2015.06.013

Baratgin, J., Over, D. E., and Politzer, G. (2013). Uncertainty and the de Finetti tables. *Think. Reason.* 16, 253–287. doi: 10.1080/13546783.2010.519564

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. (2007). The psychology of dynamic probability judgment: order effect, normative theory and experimental methodology. *Mind Soc.* 5, 53–66. doi: 10.1007/s11299-006-0025-z

Baratgin, J., and Politzer, G. (2016). "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: MIT Press), 119–142.

Beaver, D. (1992). "The kinematics of presupposition," in *Proceedings of the Eighth Amsterdam Colloquium,* eds P. Dekker and M. Stockhof (University of Amsterdam: ILLC), 17–36.

Beaver, D. (1997). "Presupposition," in *The Handbook of Logic and Language*, eds J. van Benthem and A. ter Meulen (Amsterdam: Elsevier), 939–1008.

Beaver, D., and Kramer, E. (2001). A partial account of presupposition projection. *J. Log. Lang. Inf.* 10, 147–182. doi: 10.1023/A:1008371413822

Belnap, N. D. (1970). Conditional assertion and restricted quantification. *Nôus* 1, 1–12. doi: 10.2307/2214285

Belnap, N. D. (1973). "Restricted quantification and conditional assertion," in *Truth, Modality and Syntax*, ed H. Leblanc (Amsterdam: North-Holland), 48–75.

Binet, A. (1902). *La Psychologie du Raisonnement: Recherches Expérimentales par l'Hypnotisme* [The Psychology of Reasoning: Experimental Research by Hypnosis]. Paris: Alcan.

Blamey, S. (2001). *Partial Logic*, Vol. 5. Amsterdam: Elsevier, 261–353.

Bochvar, D. A. (1938/1981). On a three-valued logical calculus and its application to the analysis of the paradoxes of the classical extended functional calculus. *Hist. Philos. Log.* 2, 87–112.

Bourne, C. (2004). Future contingents, non-contradiction, and the Law of excluded middle muddle. *Analysis* 64, 122–128. doi: 10.1093/analys/64.2.122

Braine, M., and O'Brien, D. P. (1998). *Mental Logic*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bruno, G., and Gilio, A. (1985). Confronto fra eventi condizionati di probabilità nulla nell' inferenza statistica bayesiana [Comparing conditional events of zero probability in Bayesian statistical inference]. *Riv. Mat. Sci. Econom. Soc.* 8, 141–152.

Calabrese, P. G. (1987). An algebraic synthesis of the foundations of logic and probability. *Inf. Sci.* 42, 187–237. doi: 10.1016/0020-0255(87)90023-5

Calabrese, P. G. (2002). Deduction with uncertain conditionals. *Inf. Sci.* 147, 143–191. doi: 10.1016/S0020-0255(02)00262-1

Cantwell, J. (2008). The logic of conditional negation. *Notre Dame J. Form. Log.* 49, 245–260. doi: 10.1215/00294527-2008-010

Cantwell, J. (2009). Conditionals in reasoning. *Synthese* 171, 47–75. doi: 10.1007/s11229-008-9379-6

Chrzastowski-Wachtel, P., Tyszkiewicz, J., Hoffmann, A., and Ramer, A. (2001). Definability of connectives in conditional event algebras of Schay-Adams-Calabrese and Goodman-Nguyen-Walker. *Inf. Process. Lett.* 79, 155–160. doi: 10.1016/S0020-0190(00)00219-2

Cobreros, P., Égré, P., Ripley, D., and van Rooij, R. (2014). Foreword: three-valued logics and their applications. *JANCL* 24, 1–11. doi: 10.1080/11663081.2014.909631

Coletti, G., and Scozzafava, R. (2002). *Probabilistic Logic in a Coherent Setting*. Dordrecht: Kluwer.

Cooper, W. S. (1968). The propositional logic of ordinary discourse. *Inquiry* 11, 295–320. doi: 10.1080/00201746808601531

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. (1936/1995). La logique de la probabilité. Actes du congrès international de philosophie scientifique. Sorbonne, 1935. IV: induction et probabilité, 31-39. Paris: Hermann. English translation (1995): the logic of probability. *Philos. Stud* 77, 181–190.

de Finetti, B. (1937/1964). La prevision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré, VII*, 1-67. English translation: "Foresight: its logical laws, its subjective sources," in *Studies in Subjective Probability*, eds H. E. Kyburg Jr. and H. E Smokler. (New York, NY: John Wiley), 55–118.

de Finetti, B. (1962). "Does it make sense to speak of 'good probability appraisers'?," in *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, ed I. J. Good (London: Heinemann), 357–364.

de Finetti, B. (1964/1972). Probabilità composte e teoria delle decisioni. *Rendiconti di Matematica*, 23, 128–134. English translation: "Conditional probability and decision theory," in *Probability, Induction, and Statistics. The Art of Guessing* (New York, NY: John Wiley), 13–18.

de Finetti, B. (1967). Sur quelques conventions qui semblent utiles [On some conventions that seem helpful]. *Rev. Roumaine Math. Pures Appl.* 9, 1227–1233.

de Finetti, B. (1974). *Theory of Probability*. Vol. 1 and 2. New York, NY: Wiley.

de Finetti, B. (1977/1993). "Il ruolo della probabilità nei diversi atteggiamenti del pensiero scientifico'" [The role of probability in the different attitudes of scientific thinking], in *Probabilità e Induzione; Induction and Probability,* eds P. Monari and D. Cocchi (Rome: Accademia Nazionale dei Lincei), 27–42.

de Finetti, B. (1980). *Probabilità [Probability]. Encyclopedia,* Vol. 10. Torino: Einaudi, 1146–1187.

de Finetti, B. (1995/2008). *Filosofia della probabilità* [Philosophy of probability]. Il Saggiatore: Milan. English translation: *Philosophical Lectures on Probability*. Collected, edited, and annotated by Alberto Mura (2008). Synthese Library; 340. Dordrecht: Springer.

de Finetti, B. (2006). *L'invenzione della Verità* [The Invention of Truth]. Cortina: Milan.

Delval, J. A., and Riviére, A. (1975). "Si llueve, Elisa lleva sombrero": Una investigacion psicologica sobre la tabla de verdad del condicional. ["If it rains, Elisa takes an umbrella". A psychological study on the truth table of the conditional]. *Rev. Psicol. Gen. Apl.* 136, 825–850.

Douven, I. (2016). *The Epistemology of Indicative Conditionals: Formal and Empirical Approaches*. Cambridge: Cambridge University Press.

Douven, I. (2017). "On de Finetti on iterated conditionals," in *Computational Models of Rationality*, eds C. Beierle, G. Brewka, and M. Thimm (London: College Publications), 265–279.

Douven, I., Elqayam, S., Singmann, H., and van Wijnbergen-Huitink, J. (2018). Conditionals and inferential connections: a hypothetical inferential theory. *Cogn. Psychol.* 101, 50–81. doi: 10.1016/j.cogpsych.2017.09.002

Douven, I., and Verbrugge, S. (2010). The adams family. *Cognition* 117, 302–318. doi: 10.1016/j.cognition.2010.08.015

Douven, I., and Verbrugge, S. (2013). The probabilities of conditionals revisited. *Cogn. Sci.* 37, 711–730. doi: 10.1111/cogs.12025

Dubois, D. (2008). On ignorance and contradiction considered as truth-values. *Log. J. IGPL* 16, 195–216. doi: 10.1093/jigpal/jzn003

Dubois, D., and Prade, H. (1994). Conditional objects as nonmonotonic consequence relationships. *IEEE Trans. Syst. Man. Cybern. Syst.* 24, 1724–1740. doi: 10.1109/21.328930

Dubois, D., and Prade, H. (2001). Possibility theory, probability theory and multiple-valued logics: a clarification. *Ann. Math. Artif. Intell.* 32, 35–66. doi: 10.1023/A:1016740830286

Dummet, M. (1958/1959). Truth. *Proc. Aristot. Soc New Series* 59, 141–162.

Dunn, J. M. (1975). Axiomatizing Belnap's conditional assertion. *J. Philos. Log.* 4, 343–397.

Ellis, B. (1973). The logic of subjective probability. *Br. J. Philos. Sci.* 24, 125–152. doi: 10.1093/bjps/24.2.125

Elqayam, S. (2006). The collapse illusion effect: a semantic-pragmatic illusion of truth and paradox. *Think. Reason.* 12, 144–180. doi: 10.1080/13546780500172425

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

Evans, J. St. B. T. (1972). Interpretation and matching bias in a reasoning task. *Q. J. Exp. Psychol.* 24, 193–199. doi: 10.1080/00335557243000067

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., 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., 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., Newstead, S. E., and Byrne, R. M. J. (1993). *Human Reasoning–The Psychology of Deduction*. Hove: Lawrence Erlbaum Associates.

Evans, J. St. B. T., and Over, D. E. (2004). *If*. Oxford: Oxford University Press.

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

Farrell, R. J. (1979). Implication and presupposition. *Notre Dame J. Form. Lo*g. 27, 51–61. doi: 10.1305/ndjfl/1093636522

Farrell, R. J. (1986). Material implication, confirmation, and counterfactuals. *Notre Dame J. Form. Log.* 20, 383–394. doi: 10.1305/ndjfl/1093882546

Frege, G. (1892/1952). "On sense and reference,". in *Translations from the Philosophical Writings of Gottlob Frege*, eds P. T. Geach and M. Black (Oxford: Blackwell), 56–78.

Gibbard, A. (1981). "Two recent theories of conditionals," in *Ifs: Conditionals, Belief, Decision, Chance, and Time*, eds W. L. Harper, R. Stalnaker, and G. Pearce (Dordrecht: Reidel), 211–247.

Gilio, A. (1990). Criterio di penalizzazione e condizioni di coerenza nella valutazione soggettiva della probabilità. [Penalty criterion and conditions of coherence in subjective evaluation of probability]. *Boll. Un. Mat. Ital.* 4B, 645–660.

Gilio, A., and Sanfilippo, G. (2011). "Coherent conditional probabilities and proper scoring rules," In *Proceedings of the* Seventh *International Symposium on Imprecise Probability*, (Innsbruck: ISIPTA-11), 189–198.

Gilio, A., and Sanfilippo, G. (2014). Conditional random quantities and compounds of conditionals. *Stud Log.* 102, 709–729. doi: 10.1007/s11225-013-9511-6

Goodman, I. R., Nguyen, H. T., and Walker, E. A. (1991). *Conditional Inference and Logic for Intelligent Systems- A Theory of Measure-Free Conditioning.* Amsterdam: North-Holland.

Gottwald, S. (2015). "Many-valued logic," in *The Stanford Encyclopedia of Philosophy*, ed Edward N. Zalta (Spring) Available online at: <http://plato.stanford.edu/archives/spr2015/entries/logic-manyvalued/>

Grandy, H., and Osherson, D. (2014). *Logic: A Primer for Psychologists.* Available online at: http://www.ruf.rice.edu/$\sim$rgrandy/primer.pdf and http://www.princeton.edu/$\sim$osherson/primer.pdf

Haack, S. (1974). *Deviant Logic*. Cambridge: Cambridge University Press.

Hailperin, T. (1996). *Sentential Probability Logic: Origins, Development, Current Status, and Technical Applications*. Bethlehem: Lehigh University Press.

Hailperin, T. (2011). *Logic with a Probability Semantics*. Lanham: the Rowman & Littlefield Publishing Group, Inc. and Lehigh University Press.

Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. New York, NY: Basic Books.

James, W. (1908). *Text-Book of Psychology, Briefer Course*. London: Macmillan.

Jeffrey, R. C. (1991). Matter of fact conditionals. *Proc. Aristot. Soc. Suppl. Vol.* 65, 161–183. doi: 10.1093/aristoteliansupp/65.1.161

Johnson-Laird, P. N., and Byrne, R. M. J. (1991). *Deduction*. Hove: Lawrence Erlbaum.

Johnson-Laird, P. N., and Tagart, J. (1969). How implication is understood. *Am. J. Psychol.* 82, 367–373. doi: 10.2307/1420752

Kleene, S. C. (1938). On notation for ordinal numbers. *J. Symbolic Log.* 3, 150–155. doi: 10.2307/2267778

Kneale, W., and Kneale, M. (1962). *The Development of Logic*. Oxford: Clarendon Press.

Łukasiewicz, J. (1920/1967). "On 3-valued logic," in *Polish Logic*, ed S. McCall (Oxford: Oxford University Press), 40–65.

Mandel, D. R. (2014). Do framing effects reveal irrational choice? *J. Exp. Psychol. Gen.* 143, 1185–1198.

Manktelow, K. I. (2012). *Thinking and Reasoning: Psychological Perspectives on Reason, Judgment and Decision Making*. Hove: Psychology Press.

McDermott, M. (1996). On the truth conditions of certain "if'-sentences". *Philos. Rev.* 105, 1–37.

Milne, P. (1997). Bruno de Finetti and the logic of conditional events. *Br. J. Philos. Sci 48*, 195–232. doi: 10.1093/bjps/48.2.195

Milne, P. (2004). Algebras of intervals and a logic of conditional assertions. *J. Philos. Log. 33*, 497–548. doi: 10.1023/B:LOGI.0000046072.61596.32

Milne, P. (2012). Indicative conditionals, conditional probabilities, and the "defective truth-table": a request for more experiments. *Think. Reason.* 18, 196–224. doi: 10.1080/13546783.2012.670754

Mura, A. (2009). "Probability and the logic of de Finetti's trievents," in *Bruno de Finetti Radical Probabilist*, ed M. C. Galavotti (London: College Publications), 201–242.

Mura, A. (2016). Logica dei condizionali e logica della probabilità [Conditionals logic and probability logic]. *Riv. Filos.* 57, 71–98. doi: 10.1413/82724

Muskens, R., van Benthem, J., and Visser, A. (1996). *Dynamics.* Group preprint series, Department of Philosophy, Utrecht University.

Nakamura, H., and Kawaguchi, J. (2016). People like logical truth: testing the intuitive detection of logical value in basic propositions. *PLoS ONE* 11:e0169166. doi: 10.1371/journal.pone.0169166

Nakamura, H., Shao, J., Baratgin, J., Over, D. E., Takahashi, T., and Yama, H. (2018). Understanding conditionals in the East: a Replication study of Politzer et al. (2010) with Easterners. *Front. Psychol.* 9:505. doi: 10.3389/fpsyg.2018.00505

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

Oaksford, M., and Chater, N. (2009). Précis of Bayesian rationality: the probabilistic approach to human reasoning. *Behav. Brain Sci.* 32, 69–84. doi: 10.1017/S0140525X09000284

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

O'Connor, D. J. (1951). The analysis of conditional sentences. *Mind* 60, 351–362.

Over, D. E. (2009). New paradigm psychology of reasoning. *Think. Reason.* 15, 431–438. doi: 10.1080/13546780903266188

Over, D. E. (2016). "The paradigm shift in the psychology of reasoning," in *Cognitive Unconscious and Human Rationality*, eds L. Macchi, M. Bagassi, and R. Viale (Cambridge: MIT Press), 79–97.

Over, D. E., and Baratgin, J. (2017). "The "defective" truth table: its past, present, and future," in *The Thinking Mind: A Festschrift for Ken Manktelow*, eds N. Galbraith, E. Lucas, and D. E. Over (London: Psychology Press), 15–28.

Over, D. E., and Cruz, N. (2018). Probabilistic accounts of conditional reasoning. in *International handbook of thinking and reasoning,* eds L. J. Ball and V. A. Thompson (Hove: Psychology Press), 434–450.

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. (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*, 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*, eds K. Manktelow, D. Over, and S. Elqayam (Hove,: Psychology Press), 145–166.

Politzer, G. (1981). Differences in interpretation of implication. *Am. J. Psychol.* 94, 461–477. doi: 10.2307/1422257

Politzer, G., and Baratgin, J. (2016). Deductive schemas with uncertain premises using qualitative probability expressions. *Think. Reason.* 22, 78–98. doi: 10.1080/13546783.2015.1052561

Politzer, G., Over, D. E., and Baratgin, J. (2010). Betting on conditionals. *Think. Reason.* 16, 172–197. doi: 10.1080/13546783.2010.504581

Quine, W. V. O. (1950). *Methods of Logic*. New York, NY: Holt.

Ramsey, F. P. (1926/1990). "Truth and probability," in *Philosophical Papers,* ed D. H. Mellor (Cambridge: Cambridge University Press), 52–94.

Ramsey, F. P. (1929/1990). "General propositions and causality (originally published 1929)," in *Philosophical Papers*, ed D. H. Mellor (Cambridge: Cambridge University Press), 145–163.

Reichenbach, H. (1944). *Philosophic Foundations of Quantum Mechanics*. Berkeley, CA: University of California Press.

Reichenbach, H. (1952/1953). Les fondements logiques de la mécanique des quanta [The logical foundations of quantum mechanics]. *Annales*

de l'Institut Henri Poincaré, 13, 109–158. doi: 10.1007/978-94-009-9855-1_7

Rescher, N. (1962). Quasi-truth-functional systems of propositional logic. *J. Symb. Log.* 27, 1–10. doi: 10.2307/2963674

Rescher, N. (1969). *Many-Valued Logic*. New York, NY: McGraw-Hill.

Rips, L. J. (1994). *The Psychology of Proof*. Cambridge, MA: MIT Press.

Rothschild, D. (2014). Capturing the relationship between conditionals and conditional probability with a trivalent semantics. *J. Appl. Non Class. Log*. 24, 144–152. doi: 10.1080/11663081.2014.911535

Sanfilippo, G., Pfeifer, N., Over, D. E., and Gilio, A. (2018). Probabilistic inferences from conjoined to iterated conditionals. *Int. J. Approx. Reason.* 93, 103–118. doi: 10.1016/j.ijar.2017.10.027

Schay, G. (1968). An algebra of conditional events. *J. Math. Anal. Appl.* 24, 334–344. doi: 10.1016/0022-247X(68)90035-8

Seuren, P. A. M. (1988). Presupposition and negation. *J. Semant.* 6, 175–226. doi: 10.1093/jos/6.1.175

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

Stenning, K., and Van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. Cambridge: MIT Press.

van Wijnbergen-Huitink, J., Elqayam, S., and Over, D. E. (2015). The probability of iterated conditionals. *Cogn. Sci.* 39, 788–803. doi: 10.1111/cogs.12169

Vidal, M., and Baratgin, J. (2017). A psychological study of unconnected conditionals. *J. Cogn. Psychol.* 29, 769–781. doi: 10.1080/20445911.2017.1305388

Wason, P. C. (1959). The processing of positive and negative information. *Q. J. Exp. Psychol.* 11, 92–107. doi: 10.1080/17470215908416296

Wason, P. C. (1966). "Reasoning," in *New Horizons in Psychology*, ed B. Foss (Harmondsworth: Penguin Books), 135–151.

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

HYPOTHESIS AND THEORY
published: 26 October 2018
doi: 10.3389/fpsyg.2018.02051

Imprecise Uncertain Reasoning: A Distributional Approach

*Gernot D. Kleiter**

Fachbereich Psychologie, Universität Salzburg, Salzburg, Austria

The contribution proposes to model imprecise and uncertain reasoning by a mental probability logic that is based on probability distributions. It shows how distributions are combined with logical operators and how distributions propagate in inference rules. It discusses a series of examples like the Linda task, the suppression task, Doherty's pseudodiagnosticity task, and some of the deductive reasoning tasks of Rips. It demonstrates how to update distributions by soft evidence and how to represent correlated risks. The probabilities inferred from different logical inference forms may be so similar that it will be impossible to distinguish them empirically in a psychological study. Second-order distributions allow to obtain the probability distribution of being coherent. The maximum probability of being coherent is a second-order criterion of rationality. Technically the contribution relies on beta distributions, copulas, vines, and stochastic simulation.

Keywords: uncertain reasoning, judgment under uncertainty, probability logic, imprecise probability, second-order distributions, coherence

OPEN ACCESS

Edited by:
Nathan Dieckmann,
Oregon Health & Science University,
United States

Reviewed by:
Nadia Ben Abdallah,
NATO Centre for Maritime Research and Experimentation, Italy
Edgar Merkle,
University of Missouri, United States

***Correspondence:**
Gernot D. Kleiter
gernot.kleiter@gmail.com;
gernot.kleiter@sbg.ac.at

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

Received: *12 April 2018*
Accepted: *05 October 2018*
Published: *26 October 2018*

Citation:
Kleiter GD (2018) Imprecise Uncertain Reasoning: A Distributional Approach. Front. Psychol. 9:2051. doi: 10.3389/fpsyg.2018.02051

1. INTRODUCTION

1.1. Logic, Probability, and Statistics in Models of Human Reasoning

Fifty years ago Peterson and Beach (1967) wrote a paper with the title "Man as an intuitive statistician." In the time before the heuristics-and-biases paradigm human judgments and decisions were seen on the background of Baysian statistics. In the same time human reasoning was exclusively seen on the background of classical logic. The Wason task became a prototypical experimental paradigm. One might have written a paper with the title "The human reasoner as an intuitive logician." This changed from the middle of the 1990s when probability entered the scene of human reasoning research. In 1993 *Cognition* published a special issue on the interaction between reasoning and decision making (Johnson-Laird and Shafir, 1993) with contributions, among others, by Johnson-Laird, Tversky, or Evans. Shortly afterwards Oaksford and Chater (1995) proposed to model the Wason task in terms of probabilistic information seeking. In the same year Over investigated the suppression task in terms of probabilities (Stevenson and Over, 1995). Before that time reasoning research was exclusively done on the background of logical benchmarks, while judgment under uncertainty, however, was investigated on the background of probabilistic and decision theoretic benchmarks. Reasoning investigated the human understanding of material implications (like in the Wason task), propositional inference rules (like the MODUS PONENS), inferences with quantifiers (like syllogisms), and the validity of inference forms.

The MODUS PONENS, for example, was not cast into a probabilistic format (except by George Boole more than 100 years earlier). The judgment under uncertainty community investigated updating probabilities via Bayes' theorem, calibration, and later on the heuristics and biases. Logicians had already started probability logic and default reasoning in the 1960s (Adams, 1965, 1966; Suppes, 1966)[1].

In judgment under uncertainty logical rules like the MODUS PONENS or the MODUS TOLLENS were not investigated. Inference forms of classical logic could not directly be cast into a probabilistic format. First, there was the problem of conditionals. In classical logic a conditional is a material implication. In probability logic the conditional is a conditional event to which a conditional probability may be assigned. Conditional events, however, are outside of classical logic. Second, probabilistic inference is not "truth-functional" in a way that is analog to classical logic. In classical logic the truth values of the premises determine the truth-value of the conclusion. If A is true and $A \rightarrow B$ is true, then B is true. In probability theory the probabilities of the premises of a MODUS PONENS do not exactly determine the probability of its conclusion; the premises only constrain the probability of the conclusion by lower and upper probabilities. If $P(A) = x$ and $P(B|A) = y$, then $xy \leq P(B) \leq 1 - x + xy$. Research on mental probability logic and the new (probabilistic) paradigm after the middle of the 1990s might have been published under the title "The human reasoner as an intuitive probabilist." At conferences one could follow discussions on questions like "should binary truth values be basic ingredients in models on human reasoning?"

No doubt, the adoption of probability extended and enriched the research on human reasoning. However, probability combined with some logic is still insufficient to model reasoning and decision making in a complex and uncertain environment. The reasoner as an "intuitive statistician" is missing. The intuitive statistician is required when it comes to learning, to prediction, and to decision making. A typical problem that cannot be handled in elementary probability logic but than can conveniently be handled in statistics is the *distributional precision*. By distributional precision I mean the spread-out and dispersion of a continuous distribution around a favorite value. Mental probability logic assumes precise point probabilities or probability intervals where the lower and upper bounds are again precise. Representing imprecise uncertainties by distributions opens the door to invoke an interface to frequencies observed in the outside world. We will borrow the tool of beta distributions from Bayesian statistics. Their use in psychological modeling has the advantage of providing the possibility to update beliefs in the light of new evidence and observed frequencies. "... the true power of a probabilistic representation is its ability not only to deal with *imprecise* probability assessments, but to welcome them as providing a natural basis for the system to improve with experience" (Spiegelhalter et al., 1990, p. 285). In Pfeifer and Kleiter (2006a) we used mixtures of beta distributions to model inferences with imprecise probabilities.

[1]For Adam's probabilistic validity in the more recent research on human reasoning see Kleiter (2018).

The present paper proposes first steps toward a mental probability logic based on distributions. It employs second-order probability distributions and some more recent concepts of modeling probabilistic dependence by copulas and vines. Human reasoners and decision makers should be seen as a combination of intuitive logicians, of intuitive probabilists, and of intuitive statisticians. All three levels should be addressed in the basic research questions, in the experimental paradigms, and in the normative models.

Imprecision may be expressed by various distributions. One option, for example, is the family of log-normal distributions. We made a different choice and decided for beta distributions, a family of distributions that seems to be simpler and more flexible than the log-normal. So let us, at the outset, give a short characterization of the beta family.

1.2. Beta Distribution

Throughout the contribution we will express imprecise probabilities by beta distributions. Beta distributions build a rich and flexible family of probability density functions (Johnson and Kotz, 1970; Gupta and Nadarajah, 2004). An uncertain quantity X is (standard) beta distributed in the interval $[0, 1]$ with shape parameters α and β if

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1. \tag{1}$$

For integer values the ratio of gamma functions simplifies to $(\alpha + \beta - 1)!/[(\alpha - 1)!(\beta - 1)!]$. We write for short $X \sim \mathrm{Be}(\alpha, \beta)$. The mean and the variance of the distribution are

$$\mathrm{E}(X) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{2}$$

In the present context the random variable X is a first-order probability and $p(X)$ is a second-order probability density function. In Bayesian statistics the shape parameters α and β are related to the frequencies of success and failure. α and β may be interpreted as weights of evidence, the pros and contras for a binary event, or as real or hypothetical samples sizes. $Be(1, 1)$ is the uniform distribution. If $\alpha > 1$ and $\beta > 1$ the distributions is uni-modal, if either $\alpha < 1$ or $\beta < 1$ it is J-shaped, and if $\alpha < 1$ and $\beta < 1$ it is U-shaped. **Figure 1** shows uni-modal examples.

While beta distributions do not arise exclusively in Bayesian statistics, Bayesian statistics is the field in which they are most prominent. For the assessment of subjective probability distributions Staël von Holstein proposed to fit beta distributions to quantiles 1970 and before (Staël von Holstein, 1970; Kleiter, 1981). Thomas Bayes was actually the pioneer of beta distributions in his investigation of an uncertain probability (Bayes, 1958).

The next section gives a motivating example of the application of beta distributions. Imprecision is contained in the verbal uncertainty phrases we use in everyday conversation and beta distributions may be used to represent the imprecision in a mathematical form.

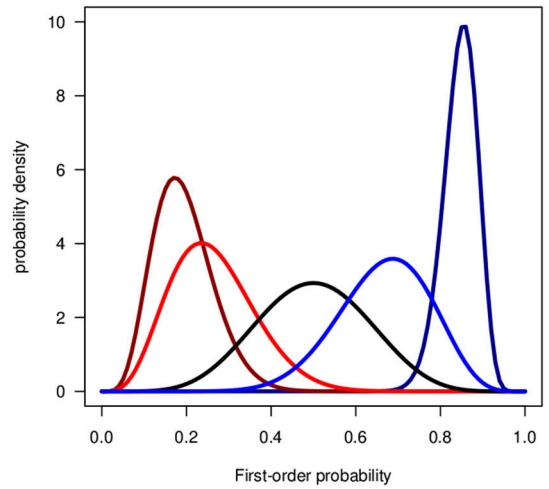


FIGURE 1 | Beta distributions for the verbal phrases of the Lichtenstein and Newman data in **Table 1**. From **(left)** to **(right)**: Very unlikely, unlikely, about as likely as not, likely, very likely.

1.3. Verbal Uncertainty Phrases

Practically all human probability judgments are imprecise. Take the following phrases in everyday communication: "very probably," "pretty sure," "highly unlikely," and so on. Verbal phrases are not only used to express degrees of belief in everyday conversation, they are also used to communicate expert knowledge, for example in geopolitical forecasting (Friedman et al., 2018) or in climate research. The Climate Science Special Report of the United States Government's (Wuebbles et al., 2017) reports a list of Key Findings. In the Climate Report each Key Finding is weighted by a verbal phrase for its likelihood. The "semantics" given to each of the phrases are shown in **Table 1**.

> "The frequency and intensity of extreme heat and heavy precipitation events are increasing in most continental regions of the world (*very high confidence*). These trends are consistent with expected physical responses to a warming climate. Climate model studies are also consistent with these trends, although models tend to underestimate the observed trends, especially for the increase in extreme precipitation events (*very high confidence* for temperature, *high confidence* for extreme precipitation). The frequency and intensity of extreme high temperature events are virtually certain to increase in the future as global temperature increases (*high confidence*). Extreme precipitation events will very likely continue to increase in frequency and intensity throughout most of the world (*high confidence*). Observed and projected trends for some other types of extreme events, such as floods, droughts, and severe storms, have more variable regional characteristics" Wuebbles et al. (2017, p. 35).

One of the first empirical studies on the interpretation of verbal uncertainty phrases in terms of numerical probabilities was performed by Lichtenstein and Newman (1967). **Table 1** shows the medians and standard deviations of the distributions of the responses of 180 persons. We represent the verbal uncertainty phrases by beta distributions. **Figure 1** shows the beta distributions fitted to the medians and standard deviations of the data.

There are two different directions in which imprecise uncertainty can be modeled, by down-shifting or by up-shifting. Down-shifting relaxes the precision of the description and works with qualitative or comparative probabilities. Baratgin et al. (2013), for example, investigated human reasoning in terms of qualitative probabilities. Up-shifting refines the level of the description on a meta-level. Describing imprecise uncertainty by distributions, as proposed in the present contribution, is an example of up-shifting.

The elementary theorems of probability theory propagate precise probabilities of the premises to precise probabilities of the conclusions. If, for example, A and B are two probabilistically independent events and $P(A) = x$ and $P(B) = y$, then $P(A \wedge B) = z = x \cdot y$. If probabilities are introduced in elementary *logical* operators or theorems, however, precise probabilities of the premises propagate to *imprecise* probabilities of the conclusions. If the two events A and B are not probablistically independent then the probability of $A \wedge B$ is an interval probability, $P(A \wedge B) = z \in [\max\{0, x + y - 1\}, \min\{x, y\}]$.

The theory of imprecise probabilities (Walley, 1991; Augustin et al., 2014) expresses imprecision by lower and upper probabilities, i.e., by *interval probabilities*. For psychological modeling, however, interval probabilities have several disadvantages. The iteration of conditional interval probabilities leads to theoretically complex solutions (Gilio and Sanfilippo, 2013). Moreover, empirically checking the endorsement of inferences may become too permissive because the responses of the participants may fall into very wide intervals. Another, more principal and theoretical difficulty poses the question how to base decisions on probability intervals. This problem was especially raised by Smets (1990) (for a review see Cuzzolin, 2012). Smets distinguished *credal* and *pignistic* degrees of belief, the first one for contemplation and the second one for action. We will tackle the question below and propose a new criterion, the maximum probability of being coherent. But let us first turn to the question of how to incorporate and propagate distributions in the framework of basic logical operators.

2. PROPAGATING IMPRECISION IN LOGICAL INFERENCES FORMS

2.1. Elementary Logical Operators

If our knowledge about the probability of an event A is represented by the beta distribution $P(A) \sim Be(\alpha, \beta)$, then our knowledge about its negation $\neg A$ should be expressed by $P(\neg A) \sim Be(\beta, \alpha)$. The parameters α and β just switch positions.

In many investigations (see for example Kleiter et al., 2002) it was observed that probability assessments of A and $\neg A$ do not add up to 1. If the participants of an experiment assess the probability of A and after a while give an assessment of $\neg A$ then usually $P(A)+P(B) \neq 1.0$. Probability judgments of "Is New York north of Rome?" and "Is Rome north of New York?" may easily

TABLE 1 | Verbal uncertainty phrases (Row A) and their numerical interpretation (Row B) as used in the US Government's climate report [Wuebbles et al. (2017, p. 35)].

A	Exceptionally unlikely	Extremely unlikely	Very unlikely	Unlikely	About as likely as not	Likely	Very likely	Extremely likely	Virtually certain
B	0–1%	0–5%	0–10%	0–33%	33–66%	66–100%	90–100%	95–100%	99–100%
C			10% (7%)	16% (10%)	50% (13%)	75% (11%)	90% (4%)		
D			Be(6, 25)	Be(5,14)	Be(7,7)	Be(12,6)	Be(66, 12)		

Row C, Medians and standard deviations (in parentheses) of the interpretation of the same verbal phrases as in Row A by 180 persons in the study of Lichtenstein and Newman (1967). Row D, Shape parameters of the fitted beta distributions shown in ***Figure 1****.*

lead to superadditivity, $P_1 + P_2 > 1$. Deviations from 1.0 may be systematic or random. Poor experimental conditions contribute to low reliability and next-best judgments. Erev et al. (1994) have shown that low reliability of probability judgments may lead to overconfidence and hyper-precision.

Let us next consider logical conjunction. For precise probabilities of the premises we have

$$\begin{aligned} &\text{If } P(A) = x \quad \text{and} \quad P(B) = y, \\ &\text{then } P(A \wedge B) = z \in [\max\{0, x + y - 1\}, \min\{x, y\}]. \end{aligned} \tag{3}$$

The lower and the upper bounds are known as the two Fréchet-Hoeffding copulas (Nelsen, 2006). Any probability assessment z in the interval is *coherent*. A probability assessment is coherent if it does not lead to a Dutch book (losing for sure). The top left panel in **Figure 2** shows lines for equal lower (upper) probabilities as functions of the marginals $P(A)$ and $P(B)$. At $(0.8, 0.6)$ the probabilities "project" to the interval $[0.4, 0.6]$.

Next we replace the precise probabilities x and y by the two random variables X and Y, where $X \sim Be(\alpha_1, \beta_1)$ and $Y \sim Be(\alpha_2, \beta_2)$. Moreover, we specify the kind and the degree of dependence between X and Y by a copula $C(x, y)$. To keep the contribution as simple as possible we will use Gaussian copulas, that is, Pearson's correlations. The coefficients will be denoted by ρ. There are many other copulas (Nelsen, 2006). The two marginal distributions of X and Y, together with the copula $C(x, y)$, determine the joint distribution with the densities $p(x, y)$ on the unit square $[0, 1]^2$. The bivariate Gaussian copula with the correlation coefficient ρ is given by

$$\begin{aligned} C(u, v) &= N_\rho(\Phi^{-1}(u), \Phi^{-1}(v)) \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} exp\left[-\frac{1}{2}\left(\frac{s^2 - 2\rho st + t^2}{1-\rho^2}\right)\right] dsdt \end{aligned} \tag{4}$$

with $s = \frac{u-\mu_u}{\sigma_u}$ and $t = \frac{v-\mu_v}{\sigma_v}$ and $\Phi^{-1}(u)$ and $\Phi^{-1}(u)$ denote the inverse of the univariate standard normal distribution function.

The unit square is analog to the 2×2 truth table in classical logic. While a truth table has only the two values 0 and 1 on its margins, the unit square has the real numbers between 0 and 1 along its two margins. In logic an operator maps the entries from the 2×2 table into $\{0, 1\}$. In the distributional approach an operator maps the densities on the unit-square to densities on the $[0, 1]$-interval. The two place operators require two mappings, one for the lower bound and one for the upper bound.

Each fixed value of the lower probability in (3) determines a contour line in the joint distribution on the unit square.

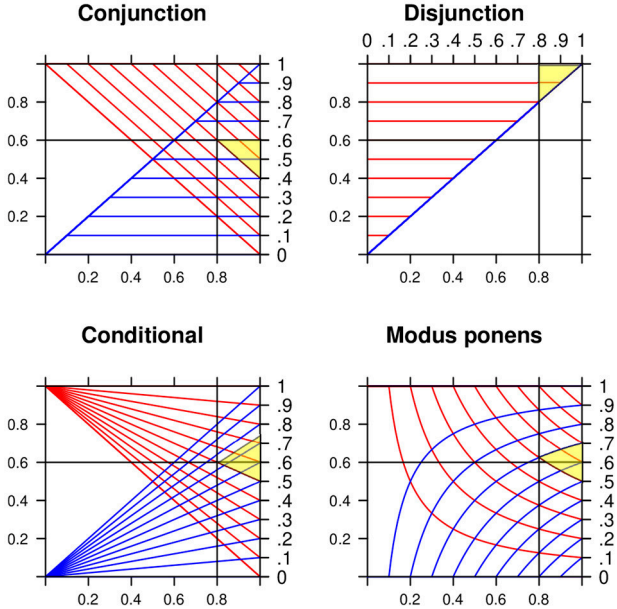


FIGURE 2 | Lower and upper probabilities for the conjunction, the disjunction, the conditional with $P(A) = X$ and $P(B) = Y$, and the MODUS PONENS with $P(A) = X$ and $P(B|A) = Y$. Numerical example for $x = 0.8$ and $y = 0.6$ (for the MODUS PONENS slightly above 0.6). The yellow shadowed areas indicate the projections to the intervals [0.4, 0.6], [0.8, 1], [0.5, 0.7], and [0.5, 0.7].

Collecting the densities along such a contour line gives the probability density for a fixed value of the lower probability. And the same holds for the upper probability. So we get two distributions, one for the lower and one for the upper probabilities. Technically in most cases these steps cannot be performed analytically in closed form. We use a stochastic simulation method implemented in the VineCopula package (Mai and Scherer, 2012; Schepsmeier et al., 2018) of the statistical software R (R Development Core Team, 2016). The R code of program for the analysis of the four inference forms discussed below is contained in the **Supplementary Material**.

We applied the stochastic simulation method to the conjunction, the disjunction, to the conditional event interpretation of the conditional (if A, then B means $B|A$) and to the exclusive disjunction. **Figure 3** shows a numerical example for each one of the four operators. The distributions of the probabilities of $X \sim Be(30, 3)$ and of $Y \sim (20, 20)$ are

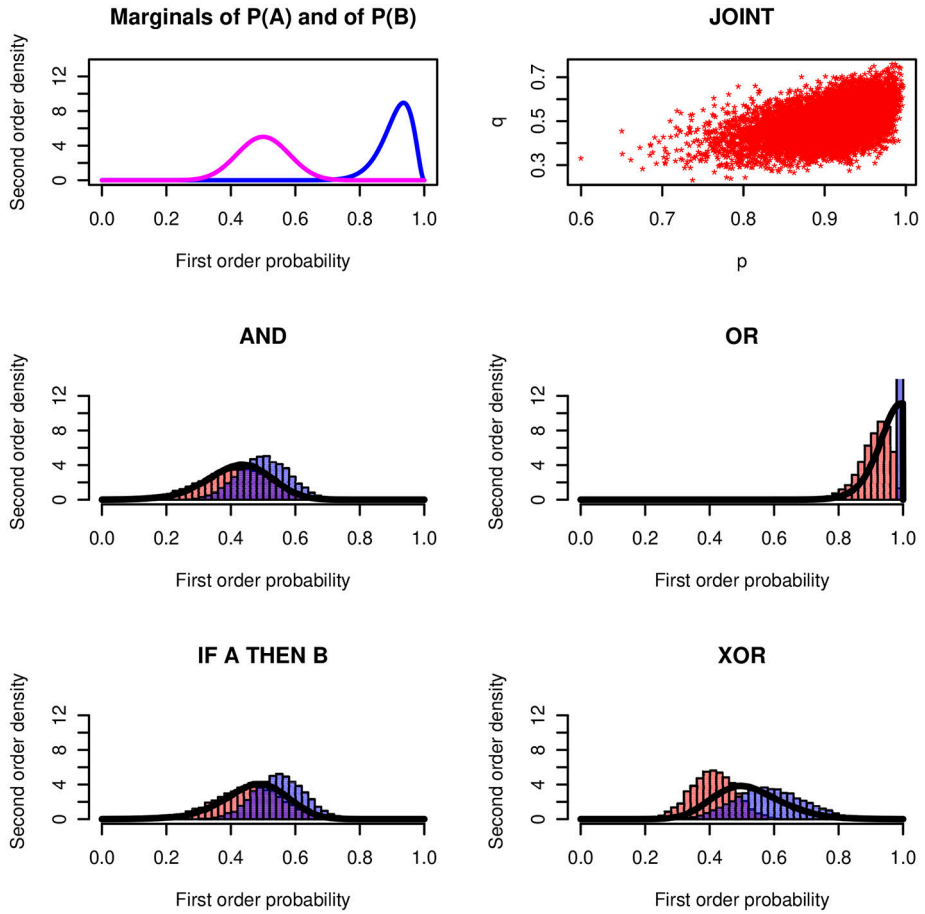


FIGURE 3 | Basic logical operators. **(Top row, Left)** Premises $P(A) \sim Be(30, 3)$ and $P(B) \sim Be(20, 20)$. **(Right)** Scatter diagram of the joint distribution with Gaussian copula $\rho = 0.5$. **(Middle and Bottom row)** Histograms of the lower and upper probabilities for AND, OR, IF-THEN, and XOR operators together with the bold lines showing the probability of being coherent. The upper probability of the disjunction degenerates at 1.

plotted in the left panel of the top row. The two first-order probabilities are correlated with the Gaussian copula $\rho = 0.5$. The scatter diagram shows the simulation of 10,000 points of the joint distribution on the unit square.

The histograms in the four panels show the relative frequencies of the lower and upper bounds resulting from the simulations. The continuous distributions approximate the probability density of being coherent. This is a meta-criterion. It corresponds to the probability that the value of a first-order probability assessment falls into the coherent interval between the two Fréchet-Hoeffding bounds. The concept will be explained below.

To consider correlations between probabilities may require a short comment. Probabilities may provide information about other probabilities. Take as an example co-morbidity in age-related diseases. Diabetes, Parkinson's and Alzheimer's disease often come together (Bellantuono, 2018). If we are 90% sure that an elderly person gets diabetes we infer that the probability that the person gets Parkinson's disease rises to a value above average. The probabilities of having the two diseases are correlated. Risks may be correlated. Assume the father of a male person suffers from prostate cancer. Knowing that the probability of having inherited some of the critical gens is high, increases the risk that the person will get prostate cancer.

Figure 3 shows a stunning result: The conjunction and the conditional (with conditional event interpretation) lead to nearly the same results[2]. It will not be possible to distinguish the two operators empirically in a psychological study. For a speaker who expresses imprecise uncertainties the *if-then* and the *and* have practically the same "meaning." This throws a new light on the conjunctive interpretation of conditionals. In Fugard et al. (2011) and Kleiter et al. (2018) we observed that about twenty percent of the participants give conjunctive interpretations of the conditional. We also observed a higher frequency of conjunctive interpretations in female participants. In real life communication, where most content is uncertain and the uncertainty is imprecise, this may not make a practical difference. We will come back to this question below after we will have introduced the distribution of being coherent.

[2]From Equation (6) is may be seen that as x approaches 1 the bounds of the conditional approach the bounds of the conjunction in Equation (3).

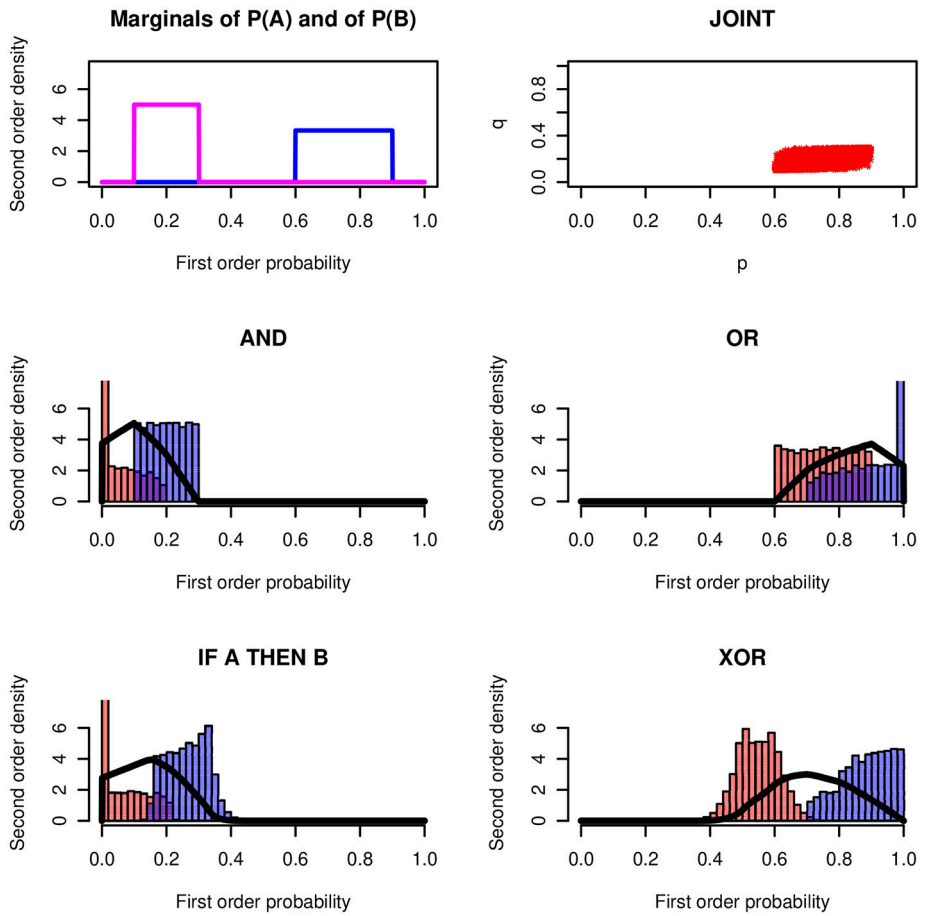


FIGURE 4 | Logical operators applied to rectangular distributions $Re(0.60, 0.90)$ and $Re(0.10, 0.30)$ and $\rho = 0.7$. The modes of the four probability-of-coherence distributions are 0.101, 0.901, 0.157, and 0.701, respectively.

Figure 4 shows the results for an example with rectangular distributions. It assumes rectangular distributions of X and Y on the intervals $Re[l_1, u_1]$ and $Re[l_2, u_2]$. Again, the conjunction and the conditional are so similar that they cannot be distinguished empirically.

Before we proceed with a discussion of the conjunction fallacy we introduce the concept of the probability of being coherent. The conjunction fallacy focuses on errors. The probability of being coherent focuses on coherent probability assessments.

2.2. The Probability of Being Coherent

Probabilistic inferences that mimic logical inferences lead from a set of precise coherent probabilities of the premises to coherent interval probabilities of the conclusion. Coherence means to not allow a Dutch book, i.e., a bet where you lose for sure[3]. Denote the inferred interval by $[w, m]$. All values between w and m are coherent.

[3]If the premises are specified by interval probabilities the situation gets more complicated and requires the concepts of g-coherence (Gilio, 1995) or the avoidance of sure loss (Walley, 1991). We do not need the concepts here.

In the present approach w and m are realizations of random variables. The probability for an assessment z to be coherent is equal to the probability that z is greater than w and less than m, i.e., $p(z \in [w, m])$. The distribution cannot be obtained in closed form. Numerical results are determined by stochastic simulation. Consider for example the conjunction of A and B with $P(A) = X \sim Be(\alpha_1, \beta_1)$, $P(B) = Y \sim Be(\alpha_2, \beta_2)$, and the copula $C(x, y)$. We perform the following steps:

1. Discretize the real numbers between 0 and 1 into n steps; we rescale the $[0, 1]$ interval by $[0, 1, \ldots, 1000]$.
2. Initialize an array $f[0], f[1], \ldots, f[n]$ of length $n + 1$ with all values equal to 0. The array will collect frequency counts.
3. Sample two random probabilities x and y from the two beta distributions of A and B; for doing this use the copula $C(x, y)$. Independence is a special case.
4. Determine the lower and upper bounds $w = \max\{0, x+y-1\}$ and $m = \min\{x, y\}$.
5. Add 1 to the frequency count of each discretized value between w and m, $f[i] = f[i] + 1$, $i = 1000 \cdot w, \ldots, 1000 \cdot m$.
6. Repeat the steps 3 to 5 N times. N may, for example, be 50,000.

7. Divide the frequency counts of the discretized values by N. The result approximates the distribution of the probability of being coherent.

We implemented these steps in R (R Development Core Team, 2016) using the package VineCopula (Schepsmeier et al., 2018). The package offers a multitude of different copulas that may be used to specify the kind and the strength of dependencies (see also Mai and Scherer, 2012).

It is rational to require that a precise probability assessment in a probabilistically imprecise world maximizes the probability of being coherent. The second-order probabilities do not lose the Dutch book criterion as claimed by Smets and Kruse (1997, p. 243). If there is a set of bets, it is reasonable to prefer that one that maximizes the probability to avoid losses. *The hierarchical construction of first- and second-order probabilities goes hand in hand with a multi-level rationality criterion.*

Smets (1990) distinguished two levels of uncertainty representation: The *credal* level—beliefs are entertained—and the *pignistic* level—beliefs are used to act. Interval probabilities are typical of the credal level. They may be entertained in the cognitive representation of uncertainty. Practical decisions, however, require the selection of precise point values that maximize, e.g., expected utility. Smets' pignistic probabilities are different from the maximum probability of being coherent. We note that point probabilities are not always required for decision making. In decision theory, economics, and risk management *distributions* and not only exact probabilities are compared. The criterion of stochastic dominance (Sriboonchitta et al., 2010) may, for example, be applied to two distributions of being coherent.

The discriminatory sensitivity of the logical connectives may be studied by measuring the distance between two distributions of being coherent. A well known measure for the distance between two distributions is the Kullback-Leibler distance. Because of the stochastic simulation the distributions of the probability of being coherent are discrete, in our case having $N = 1,000$ increments. The Kullback-Leibler distance between two probability distribution P and Q is given by

$$D(P; Q) = \sum_{i=1}^{N} P(x_i) \log \frac{P(x_i)}{Q(x_i)}, \quad \text{where } x_1 = 1/N, x_2 = 2/N, \ldots, x_N = 1. \tag{5}$$

Numerical probabilities equal to zero were set equal to 0.0001. **Table 2** shows the distances between ten pairs of distributions, three kinds of beta distributions, and the two correlation coefficients $\rho = 0.5$ and $\rho = -0.5$.

The left side of **Table 2** contains distances from the uniform distribution (UFD). These distances are all high and relative insensitive to the kind of the distributions of $P(A)$ and $P(B)$ and the correlation coefficients ρ. The greatest distances are between OR and UFD and between AND and UFD.

On the right side of **Table 2** small distance indicate that the probabilistic semantics of the two operators is similar. The smallest value of $D(P; Q) = 0.14$ is obtained for the distance between IF and AND for $P(A) \sim Be(30, 3)$ and $P(B) \sim (20, 20)$, that is, for one distribution with a high mean of 0.91 and one distribution with a mean of 0.5. This may be related to the empirical finding that about twenty percent of the interpretations of if-then sentences are conjunction interpretations (Fugard et al., 2011; Kleiter et al., 2018).

The conclusion that may be drawn from this analysis is: *The difference or the similarity of the probabilistic meaning of two logical operators depends on the high, middle, or low probabilities of the events and on the copula between the two.* This makes the empirical investigation of the semantics of the logical operators in reasoning and everyday language more difficult than often assumed. This holds, for example, for our own experiments where we used truth-table tasks in which relative frequencies were selected that may discriminate conjunctions, disjunctions, conditionals etc. This is only possible if the frequencies presented to the participants in the truth tables are close to being equally distributed and not rather high or low.

We next turn to the conjunction fallacy, one of the best known fallacies in the heuristics and biases paradigm. We will see that imprecision is a factor that may explain the fallacy at least to some degree.

2.3. Conjunction Fallacy

In the same way as we asked for the probability of being coherent, we may ask for the probability of being incoherent. A prototypical example for incoherent probability judgments is the Linda task (Tversky and Kahneman, 1983):

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was especially concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Rank order the probabilities for
>
> - Linda is a bank teller.
> - Linda is active in the feminist movement.
> - Linda is a bank teller and is active in the feminist movement.

Many people think the conjunction is more probable than one or even both its conjuncts. They are victims of the conjunction fallacy.

Like many other tasks in the literature on fallacies and biases, the Linda task is an example for highly imprecise probabilities. Denote "Linda is a bank teller" by A, "Linda is a feminist" by B and assume $P(A) = X \sim Be(\alpha_1, \beta_1)$, $P(B) = Y \sim Be(\alpha_2, \beta_2)$, and a Gaussian copula with $\rho = 0.7$.

You create two vague ideas of the probabilities of A and B, modeled here by two beta distributions. Next you think about reasonable values for the probabilities of the conjunction, modeled here by the distribution of the probability of being coherent. In the terminology of Smets the three distributions belong to the *credal* level. The beliefs are just "entertained" and their imprecision is part of their representation. When it is time for judgment one value x is sampled from the distribution for A and one value y from the distribution for B. Now if you really think hard you *infer* the third value z on the basis of x and y and

TABLE 2 | Logical operators: Kullback-Leibler distances between the second order distributions of the probability of being coherent and the uniform distribution (UFD) and between the distributions of the conjunction (AND), the disjunction (OR), the conditional (IF) and the exclusive disjunctions (XOR).

P(A)	*P(B)*	ρ	AND UFD	OR UFD	IF UFD	XOR UFD	OR AND	IF AND	XOR AND	IF OR	XOR OR	XOR IF
Be(30,3)	Be(20,20)	0.5	7.80	8.78	7.80	7.74	11,06	0.14	0.63	9.94	9.30	0.21
Be(30,3)	Be(20,20)	−0.5	8.77	7.80	7.80	7.74	11,06	0.47	0.19	9.40	9.77	0.21
Be(100,10)	Be(20,2)	0.5	8.22	9.10	8.45	8.17	6.78	3.16	10.43	0.716	10.47	10.46
Be(100,10)	Be(20,2)	−0.5	8.55	9.34	8.54	8.33	8.86	4.80	10.63	1.35	10.63	10.63
Be(20,100)	Be(5,20)	0.5	8.55	7.88	6.91	7.67	6.41	6.18	3.25	3.13	1.46	0.68
Be(20,100)	Be(5,20)	−0.5	8.66	8,12	6.92	7.77	8.73	6.51	4.82	4.16	1.98	0.74

ρ denotes the value of the Gaussian copula.

TABLE 3 | Probability of a conjunction error.

Beta distribution	Be(1,1)	Be(2,2)	Be(4,2)	Be(8,2)	Be(16,2)
Probability of a conjunction error	0.50	0.50	0.33	0.22	0.15

The beta distribution of one conjunct is held constant at Be$(30, 5)$*; Gaussian copula* $\rho = 0.5$.

the inferred value may be coherent. If you are lazy you sample a third time, now a value z from the distribution for being coherent. You come up with a judgment z that is *decoupled* from x and y. If you think hard your judgment of z is coupled to the precise values x and y, with less strain it is sampled from a distribution. In this case z may easily exceed the upper bound of the conjunction probability, i.e., the minimum of x and y and the result is a conjunction error. The probability of this one-sided incoherence corresponds to the probability that z is in the interval between the upper bound m and 1, $P(z \in [m, 1])$.

Applying simulation methods again gives a surprising result. If my probability assessment of "Linda is a bank teller" is close to 0.5 or if my assessment of "Linda is active in the feminist movement" is close to 0.5, the probability of a conjunction error may be as high as 50%. *Imprecise probabilities may induce a high percentage of conjunction errors*. If the location of the central tendency of one of the marginals is close to 0.5, then the probability of a conjunction error is close to 0.5. The probability decreases when both means move away from 0.5. The size of the correlation (or the copula parameter) does nearly not matter. **Table 3** gives a few numerical examples.

We next turn to uncertain conditionals, the salt in the soup of probability logic. The interpretation of conditionals by humans was and is an especially important topic in human reasoning research. Imprecise conditionals were studied in terms of lower and upper probabilities. In the next section we will turn to distributional imprecision.

2.4. Conditional

Modeling conditioning with imprecise probabilities is an intricate problem. This is seen from the many different proposals made in many-valued logic, in work on lower probabilities and the Dempster-Shafer belief functions, or in work on possibilistic and fuzzy approaches. In the coherence approach inferences where the *conclusion* is a conditional require special methods. The extension of the Fundamental Theorem of de Finetti to conditional probabilities is due to Lad (1996). He also explains how numerical results are found by linear in-equalities and fractional programming (Lad, 1996).

The psychological literature reports many experiments on the interpretation of uncertain conditionals.The *truth table method* is used to distinguish between the material implication of classical logic and the conditional event interpretation. Especially the "new probabilistic paradigm" (Over, 2009; Elqayam, 2017) in reasoning research has used this task. The task is based on the truth values of the antecedent and the consequent. I, the experimenter, show you, the participant, the four combinations of the binary truth values of A and of B together with their associated probabilities. You tell me the probability you assign to "If A then B." I infer on which truth values you were attending and this allows me to reconstruct your logical interpretation of the conditional.

Given $P(A) = x$ and $P(B) = y$ the probability of $P(B|A) = z$ is in the interval

$$z \in \left[\max\left\{0, \frac{x+y-1}{x}\right\}, \min\left\{1, \frac{y}{x}\right\}\right], \quad x > 0. \tag{6}$$

The **Figures 3**, **4** show examples for the distribution of $P(B|A)$, the probability of a conditional. We have already pointed out that the results for the conjunction and the conditional can be very similar.

For the material implication (denoted by $A \rightarrow B$) this is different. Given $P(A) = x$ and $P(B) = y$ the probability of $P(A \rightarrow B) = z$ is in the interval

$$z \in [1 - \min\{y, 1 - x\}, \min\{1 - y + x, 1\}]. \tag{7}$$

The lower and upper probabilities are equivalent to those of the disjunction $\neg A \vee B$. If the probability of the antecedent $P(A)$ is high then the distribution of the lower and upper probabilities and the probability of being coherent are very similar to the disjunction $A \vee B$. With increasing $P(A)$ the distributions of $\neg A \vee B$ and $A \vee B$ get more and more indistinguishable. In

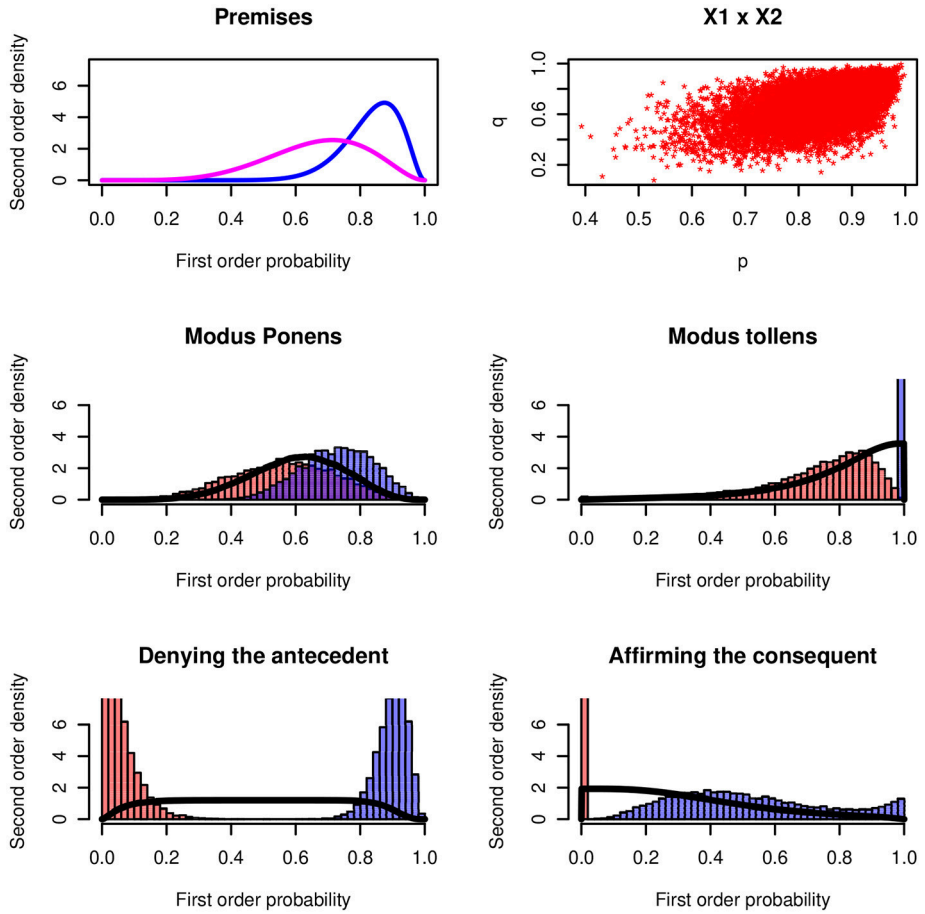


FIGURE 5 | Four inferences rules. **(Upper panels)** Probability distribution of the minor premise and the major premises $P(B|A)$. Histograms of the lower and upper probabilities of the four rules. The continuous distributions show the distributions of the probability of being coherent.

an imprecise probabilistic environment the question "material implication or disjunction?" does not matter. The question "conditional event or material implication?", however, makes a big difference: The conditional event interpretation leads to much lower probabilities than the material implication. This is a highly relevant aspect for the interpretation of *if-then* sentences in the context of risk assessment.

The interpretation of conditionals leads us to the next section, to logical inference rules. Psychologists have often investigated the MODUS PONENS along with the MODUS TOLLENS and two logically non-valid argument forms.

2.5. The MP-quartet

Four inference rules were often investigated in the psychology of human reasoning: The quartet of the MODUS PONENS, the MODUS TOLLENS (both logically valid) and the argument forms of DENYING THE ANTECEDENT and AFFIRMING THE CONSEQUENT (both logically nonvalid), here called "the MP-quartet" for short. The MODUS PONENS

$$\text{From } \{\text{if } A \text{ then } B,\ A\} \text{ infer } B$$

is the best known and most important inference rule in deductive logic. It is endorsed by practically all people (Rips, 1994). If the premises are uncertain and the conditional is interpreted as a conditional event we have in terms of point probability:

$$\text{From } \{P(B|A) = x, P(A) = y\} \text{ infer } P(B) = z, \text{ and } z \in [xy, 1 - y + xy]. \tag{8}$$

For the lower and upper bounds for the three other rules see for example (Pfeifer and Kleiter, 2005).

Figure 5 shows the results for the four inference rules for a numerical example. The premises have the distributions $X \sim Be(15, 3)$, $Y \sim Be(6, 3)$, and the Gaussian copula $\rho = 0.5$[4].

The MODUS PONENS has a maximum probability of being coherent that is close to the distribution of the minor premise $P(A)$. For the MODUS TOLLENS the maximum probability is at 1.0. The MODUS TOLLENS is the strongest inference rule (Pfeifer and Kleiter, 2005, 2006b). Psychologically the MODUS TOLLENS is difficult and complex; it's a "backwards" rule and it involves two

[4]Denying the antecedent and affirming the consequent degenerate at 0; the MODUS TOLLENS degenerates at 1.

negations. Usually the endorsement is much lower than for the MODUS PONENS.

The two logically non-valid inference forms lead to probabilities of being coherence that are close to uniform distributions. In a psychological investigation the two rules should stick out by the *variance* of the probability judgments. More or less any probability judgment in $[0, 1]$ is coherent.

The following section applies distributional imprecision to a series of examples. Most of them are well-known from the psychological literature but the inclusion of imprecision into their analysis leads to new properties and results.

3. APPLICATIONS AND EXAMPLES

3.1. Natural Sampling

One of the best known fallacies in judgment under uncertainty is the base rate neglect (Kahneman and Tversky, 1973; Bar-Hillel, 1980; Koehler, 1996). A doctor may, for example, neglect the prevalence of a disease and concentrate only on the likelihood of a symptom given the disease. While this is often a major fallacy, there are situations in which base rate neglect is completely rational. This holds also for beta distributions: Assume the shape parameters α and β of a distribution $Be(\alpha, \beta)$ are equal to the frequency of a binary feature in a sample of n observations, $n = \alpha + \beta$. Split the total sample into two subsamples so that the sample sizes add-up to n. So the subsample sizes are not pre-planned. In statistics this is called *natural sampling* (Aitchison and Dunsmore, 1975). We have $Be(\alpha_1, \beta_1)$, $Be(\alpha_2, \beta_2)$ and $\alpha = \alpha_1 + \alpha_2$ and $\beta = \beta_1 + \beta_2$, and $n = \alpha_1 + \alpha_2 + \beta_1 + \beta_2$. For natural sampling it was proven (Kleiter, 1994) that the base rates in Bayes' Theorem are "redundant" and may be ignored. The result for precise probabilities has often been used by Gigerenzer within his frequentistic approach (Gigerenzer and Hoffrage, 1995; Kleiter, 1996).

Ignoring base rates may not only be rational for precise but also for imprecise probabilities. For natural sampling it holds that if the knowledge about the prevalence of a disease H is represented by the beta $P(H) \sim Be(\alpha, \beta)$ and the conditional probabilities of a symptom D are represented by the betas $P(D|H) \sim Be(\alpha_1, \beta_1)$ and $P(D|\neg H) \sim Be(a_2, b_2)$, then the posterior distribution of the disease given the symptom D is simply

$$\begin{gathered} P(H|D) \sim Be(\alpha_1, \alpha_2), \\ \text{mean} = \tfrac{\alpha_1}{\alpha_1+\alpha_2}, \ \text{variance} = \tfrac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}. \end{gathered} \tag{9}$$

If frequencies are used to update subjective probabilities and if (and only if) natural sampling conditions hold, the resulting degrees of belief remain in the family of beta distributions, i.e., the distributions are natural-conjugates. Note that (relative) frequencies and probabilities are not the same. The frequencies are used to estimate probabilities and the representation of the imprecision of these estimates is an integral part of any statistical approach. The property of natural sampling extends to multivariate Dirichlet distributions and is thus helpful to represent imprecise degrees of belief in more complex environments. If the natural sampling assumption is dropped, then vines and copulas offer elegant methods to model the representation and propagation of degrees of belief.

3.2. Rips Inference Tasks

To show that a wide range of logical inference tasks can be modeled within the distributional approach we discuss very briefly two examples from Rips (1994). Rips compared the predictions of his proof-logical PSYCOP model with empirical data. He investigated 32 inference problems of classical sentential logic. Among them the following one:

> IF Betty is in Little Rock THEN Ellen is in Hammond. Phoebe is in Tucson AND Sandra is in Memphis. Is the following conclusion true: IF Betty is in Little Rock THEN (Ellen is in Hammond AND Sandra is in Memphis) (Rips, 1994, p. 105).

When we represent the conditional by a conditional event[5] and first introduce precise probabilities:

$$\frac{\begin{array}{l} P(B|A) = x \\ P(C \wedge D) = y \end{array}}{P(B \wedge D|A) \in [0, x]}$$

The interval probability of the conclusion, $P(B \wedge D|A) \in [0, x]$, is easily obtained after seeing that the probability of the conjunctive premise is irrelevant. $P(D)$ is greater than $P(C \wedge D)$ and may maximally be 1. The upper probability of the conclusion is thus $P(B \wedge D|A) = \frac{P(A \wedge B \wedge D)}{P(A)}$ and $P(B \wedge D|A) = P(D)\frac{P(A \wedge B)}{P(A)} = P(B|A)$. Analog relationships hold for the probability distributions.

In a second step beta distributions for the premises are introduced, say X and Y, and by stochastic simulation the distributions for the lower and upper probabilities and the distribution of the probability of being coherent are determined. The distribution of the probability of being coherent is practically uniform over the range between 0 and the mean of X. For high probabilities of the conditional premise the inference is inconclusive. In classical logic and in the proof-logical approach of Rips the inference is valid.

Here is a second example (Example M in Rips, 1994, p. 151):

$$\frac{\begin{array}{l} \neg A \\ B \end{array}}{\neg(A \wedge C) \wedge (B \vee D)}$$

With $P(\neg A) = x$ and $P(B) = y$ the probability of the conclusion is in the interval $z \in [\max\{0, x+y-1\}, 1\}]$. The lower probability is the same as the lower probability of a conjunction. If x and y are less than 0.5, then the inference is noninformative and the distribution of the probability of being coherent is a uniform distribution. The inference was endorsed by only 22.2% of the participants.

[5]Note that Rips (1994, p. 125) prefers the suppositional interpretation of the conditional; the domain of a conditional consists only of those possibilities in which the antecedent is true. PSYCOP rejects the paradoxes of the material implication!

We next turn to an example from the judgment under uncertainty domain. It may be considered as an example of Ockham's razor (Tweney et al., 2010) where less is more.

3.3. The Doherty Task

For the conjunction of n events we have: If $P(D_i) = \alpha_i$ for $i = 1, \ldots, n$, then

$$P(D_1 \wedge D_2 \wedge \ldots \wedge D_n) \in \left[\max \left\{ \sum_{i=1}^{n} \alpha_i - (n-1)\}, \min\{\alpha_i\} \right\} \right]. \tag{10}$$

This is a straightforward generalization of the elementary conjunction rule. Such generalizations were first investigated by Gilio (2012) and are also studied in Wallmann and Kleiter (2012a,b, 2014a,b). There is a psychologically interesting property of such generalizations. It is the phenomenon called *degradation*. As n, the number of events in the generalization, increases the inferences become less and less informative. More information leads to less conclusive inferences.

An example in the field of judgment under uncertainty is the so called pseudodiagnosticity task introduced by Michael Doherty (Doherty et al., 1979, 1996; Tweney et al., 2010; Kleiter, 2013). It was analyzed with second-order distribution by Kleiter (2015).

> Assume you are a physician and you are 50% sure that one of your patients is suffering from disease H, $P(H) = 0.5$. You know that the probability that if the patient is suffering from H, the patient shows symptom D_1 is 0.7, $P(D_1|H) = 0.7$. You may obtain just one more piece of information. There are three options:
>
> 1. $P(D_2|H)$, the probability of a second symptom given the presence of the disease,
> 2. $P(D_1|\neg H)$, the probability of the first symptom given the absence of the disease, or
> 3. $P(D_2|\neg H)$, the probability of the second symptom given the absence of the disease.
>
> What is your choice?

Most people select $P(D_2|H)$. Actually $P(D_1|\neg H)$ is the best choice. With $P(D_1|\neg H)$ Bayes' theorem gives the posterior probability

$$P(H|D_1) = \frac{P(H)P(D_1|H)}{P(H)P(D_1|H) + [1 - P(H)]P(D_1|\neg H)}. \tag{11}$$

Before any of the three options is selected, the posterior probability is in the interval (Tweney et al., 2010)

$$P(H|D_1) \in \left[\frac{P(H)P(D_1|H)}{P(H)P(D_1|H) + 1 - P(H)}, 1 \right]. \tag{12}$$

If however, as most participants do, $P(D_2|H)$ is selected, then the interval is

$$P(H|D_1, D_2) \in \left[\frac{P(H)P(D_1, D_2|H)}{P(H)P(D_1, D_2|H) + 1 - P(H)}, 1 \right]. \tag{13}$$

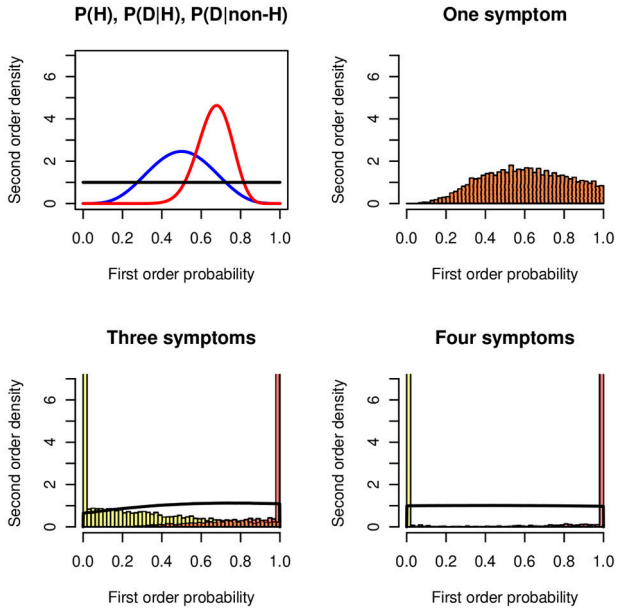


FIGURE 6 | Degradation in the Doherty task. **(Top left panel)** The symmetric beta distributions $Be(5, 5)$ of $P(H)$ (blue), $Be(20, 10)$ for $P(D|H)$ (red), and the uniform distribution $Be(1, 1)$ for $P(D|\neg H)$ (black). **(Top right panel)** Second-order posterior distribution of the probability of H when the distributions of the likelihoods $P(D|H)$ and $P(D|\neg H)$ are known. **(Bottom panels)** Lower and upper distributions of the probability of H when the distributions of the likelihoods of three **(Left panel)** and four **(Right panel)** symptoms are known; all likelihood distributions are $Be(20, 10)$ and $Be(1, 1)$, respectively. The black line shows the probability of being coherent.

The interval in (13) is wider than the interval in (12) as

$$P(D_1, D_2|H) \leq \min\{P(D_1|H), P(D_2|H)\} \leq P(D_1|H).$$

Selecting $P(D_1|\neg H)$ results in a precise point probability while selecting $P(D_2|H)$ results in an interval that is *wider* than the initial one.

If we continue to select only the "affirmative " likelihoods given H and not those given $\neg H$, then the intervals get wider and wider and after a few more steps become noninformative, that is, $[0, 1]$. The additional information imports noise. **Figure 6** shows an example for $P(H) \sim Be(5, 5)$, $P(D_i|H) \sim Be(20, 10)$, and $P(D_i|\neg H) \sim Be(1, 1)$. For $i = 1$ there is one posterior distribution, the lower and the upper distributions coincide; for $i = 3$ and $i = 4$ the lower and upper distributions get close to 0 and 1. The probability of being coherent becomes a uniform distribution. One reason that contributes to the degradation effect are the unknown probabilities of the conjunctions $P(D_1|H) \wedge \ldots \wedge P(D_n|H)$ and $P(D_1|\neg H) \wedge \ldots \wedge P(D_n|\neg H)$.

The Doherty task demonstrates that we should compare the results from experimental groups with those from control groups. The preference for selecting the affirmative likelihood only is seen as a confirmation bias: people do not consider alternative hypotheses. The phenomenon that more information may induce

more imprecision has been studied in Wallmann and Kleiter (2012a,b, 2014a,b) and Kleiter (2013).

Technically the analysis of a *multivariate* problem like the Doherty task requires stochastic simulation in *vines*. "Vines are graphical structures that represent joint probabilistic distributions. They were named for their close visual resemblance to grapes ..." Kurowicka and Joe (2011, p. 1). Vines may be compared to Bayesian networks. In psychology Bayesian networks were used, for example, to model uncertain reasoning (Oaksford and Chater, 2007), to model causal reasoning (Tenenbaum et al., 2007), word learning (Xu and Tenenbaum, 2007), or to model cognitive development (Gopnik and Tenenbaum, 2007). Bayesian networks encode conditional independencies and represent the (usually precise) joint probabilities in tables. Vines encode marginal probabilities and (partial) correlations, or more generally, copulas. Psychologically it is more plausible that humans encode multivariate uncertain structures by their (conditional) dependencies and not by their (conditional) independencies. Moreover, encoding marginal probabilities is much easier than encoding multivariate probability tables. There is no space here for further speculations. For the mathematical treatment of vines the reader is referred to Kurowicka and Cooke (2004, 2006), Kurowicka and Joe (2011), and Mai and Scherer (2012).

A psychologically interesting difference between Bayesian networks and vines is that vines encode dependencies "directly" by (partial) correlations (actually copulas) and not by conditional probabilities. It is highly plausible (but seldom investigated) that humans encode the strength of a dependence not by a probability table but by a one-dimensional quantity.

While Bayesian networks rely on (conditional) independence assumptions, vines rely on *copulas*. Copulas encode dependencies. To keep the present text simple we use Gaussian copulas (correlations) only (see Equation 4). The recent advances in the theory of copulas and vines, and the development of software for the *simulation* methods allow to model multivariate imprecise inference. There is not enough space here to discuss a more complex example, but see the study of the Doherty's pseudodiagnosticity task in (Kleiter, 2015). The suppression task in the following section involves three variables.

3.4. Suppression Task

The Suppression Task was introduced by Byrne (1989). She observed that while a simple MODUS PONENS is endorsed by nearly all people, the endorsement decreases substantially when an *additional* conditional premise is introduced. The additional premise *suppresses* the acceptance of the conclusion. **Table 4** shows Byrne's by now classical example: The simple MODUS PONENS "from $\{P1, P3\}$ infer C" is endorsed by 96% of the participants in Byrne's Experiment 1. When the additional premise P2a is included, "from $\{P1, P2a, P3\}$ infer C" the endorsement drops to 38%. When the alternative premise $P2b$ is introduced, "from $\{P1, P2b, P3\}$ infer C," the endorsement is the same as for the simple MODUS PONENS.

In an abstract formal system the second premise is logically and probabilistically irrelevant. It has no impact upon the conclusion, neither upon its truth nor upon its probability.

TABLE 4 | The various premises and the conclusion in the Suppression Task.

*P*1	Main conditional	If Mary has an essay to write, then she will study late in the library.
P2a	Additional conditional	If the library is open, then she will study late in the library.
P2b	Alternative conditional	If Mary has some textbook to read, then she will study late in the library.
*P*3	Categorical premise	Mary has an essay to write.
C	Conclusion	Mary will study late in the library.

Attending to the semantic content of the conditional premises, however, leads to a reinterpretation of the inferences. The conditionals $P1$ and $P2$ have the same consequent and Mary can only study late in the library if the library is open. Thus for the additional conditional the semantic content (Byrne, 1989) invites a conjunctive interpretation of the antecedent, {if $A \wedge B$ then C, A}. The alternative conditional P2b, however, invites a disjunctive interpretation of the antecedent, {if $A \vee B$ then C, A}.

The distributional interpretation of the three different inferences are:

1. Simple MODUS PONENS: $P(C|A) = X, P(A) = Y$.
2. Conjunctive antecedent: $P(C|A \wedge B) = X$, $P(A \wedge B) = Y$. We note that if $P(A) = x$ and $P(B)$ is unknown and thus may have any value between 0 and 1, $P(A \wedge B)$ is in the interval $[0, x]$. The bounds for the MODUS PONENS are $z \in [0, 1 - x + xy]$
3. Disjunctive antecedent: $P(C|A \vee B) = X$, $P(A \vee B) = Y$. $P(B)$ is unknown and $P(A \vee B)$ may have any value in the interval $[x, 1]$. The bounds for the MODUS POENS are $z \in [xy, y]$.

Figure 7 shows the distributions of the lower and the upper bounds and of the probability of being coherent. The example uses the following distributions: (1) For the simple MODUS PONENS $P(A) = X \sim Be(10, 5)$ and $P(C|A) = Y \sim Be(20, 5)$. (2) For the conjunctive interpretation (additional conditional) $P(A \wedge B) = X \sim Be(10, 5)$ and $P(C|A \wedge B) = Y \sim Be(20, 5)$ (3) For the disjunctive interpretation (alternative conditional) $P(A \vee B) = X \sim Be(10, 5)$ and $P(C|A \vee B) = Y \sim Be(20, 5)$.

In the figure the simple MODUS PONENS and the disjunctive antecedent (If Mary has an essay to write or if Mary has a textbook to read) lead to very similar results. The conjunctive antecedent (If Mary has an essay to write and if the library is open) leads to a very flat distribution. The distribution of the lower bound is degenerate at zero. The probability of the conjunction is much lower than the probability of the disjunction.

The distributional approach models the results of the Suppression Task pretty well. Moreover, it provides quantitative predictions for the differences in the various experimental conditions.

The suppositional interpretation of an "if H then E" sentence assumes H to be true. Also in a conditional probability $P(E|H)$ the event H is assumed to be true. Jeffrey pointed at cases where observations are blurred. Under candle light the color of an object may be ambiguous. How to condition on soft evidence? Jeffrey was the pioneer of the analysis of soft evidence to which we will turn next.

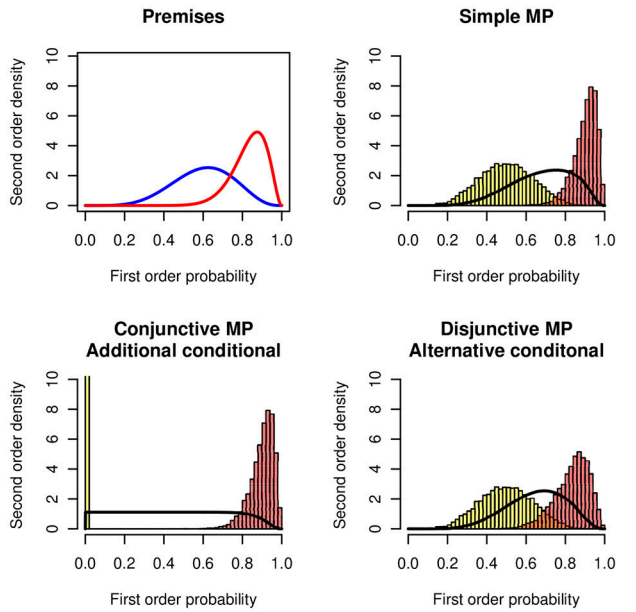


FIGURE 7 | MODUS PONENS in the Suppression Task. **(Top panels)** Probability distributions of the premises $P(C|A) \sim Be(15, 3)$ **(Right)** and $P(A) \sim Be(6, 4)$ **(Left)**. Simple MODUS PONENS: Lower and upper histograms of the probability of the conclusion $P(C)$. **(Bottom panels, Left)** The premises are interpreted as a conjunction, $P(C|A \wedge B)$ and $P(A \wedge B)$. **(Right)** The premises are interpreted as a disjunction , $P(C|A \vee B)$ and $P(A \vee B)$.

3.5. Soft Evidence

Usually *conditioning* updates probabilities in the light of *hard* evidence, that is, the conditioning event is supposed to be *true*. But what if the conditioning event is only uncertain? Jeffrey introduced "Jeffrey's rule," a proposal of how to update probabilities by *soft* evidence (Jeffrey, 1965, 1992, 2004). Historically the problem was ready posed by Donkin (1851) and his solution is equivalent to Jeffrey's rule (for a proof see Draheim, 2017). Draheim gives an overview of the literature in Appendix A of his monograph. Jeffrey's rule has been criticized by several authors (Levi, 1967; Diaconis and Zabell, 1982; Wedlin, 1996; Halpern, 2003; Jaynes, 2003). The rule is *non-commutative*, i.e., it is not invariant with respect to the order of updating. Moreover, it involves an independence assumption. For a psychological investigation of Jeffrey's rule see Hadjichristidis et al. (2014).

In the present approach it is straightforward to update probabilities by evidence that is probable only. We have two random variables X and Y (first-order probabilities). We want to know the (second-order) distribution of Y given a fixed value of X. The problem is analog to a regression problem in statistics: The distribution of Y is predicted on the basis of a given value of $X = x$. The distributional approach offers a direct representation of Jeffrey's problem.

Figure 8 shows a numerical example. On the left side the unit square $[0, 1]^2$ and the contour lines from the bivariate joint distribution resulting from two beta marginals and a Spearman copula[6]. On the right side the two marginals and the distribution of Y at $X = 0.9$. The contour lines and the distribution at the cutting point 0.9 is obtained by stochastic simulation.

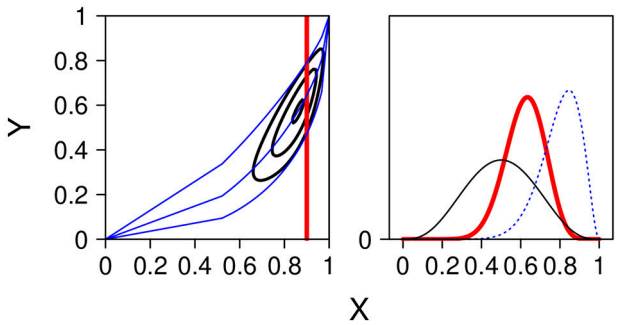


FIGURE 8 | (Left panel) Contour lines of the joint distribution with the marginals $X \sim Be(9, 3)$ and $Y \sim Be(4, 4)$, and Spearman correlation $\tau = 0.5$. Regression line at $x = .9$ (quantile at 0.5) together with 90 % confidence band (quantile at 0.05 and 0.95). **(Right panel)** The two marginal betas and the conditional distribution $p(y|x_0 = 0.9)$ along the vertical line in the contour plot.

4. DISCUSSION

We have distinguished logical, probabilistic, and statistical principles and argued that for a plausible model of human reasoning ingredients from all the three domains are necessary. We have seen that the constraints of probability logic induce only lower and upper probabilities, or lower and upper distributions in the case of imprecision; they do not lead to exact point probabilities, or to just one distribution in the case of imprecision. To overcome this kind of indeterminacy we have introduced the concept of the *probability of being coherent*. One may follow the proposal of Smets (1990) and distinguish *credal* and *pignistic* degrees of belief, corresponding to the whole distribution for the cognitive representation and the maximum for selecting just one favorite value. It is rational to base one's decisions on values obtaining a maximum probability of being coherent.

We have investigated the differences between the logical conjunction and the conditional. For not too extreme probabilities these differences may be small, so small that it will be impossible to distinguish the two interpretations empirically. We observed that in typical truth table tasks about twenty percent of the participants interpret if-then sentences as conjunctions (Fugard et al., 2011; Kleiter et al., 2018). In the context of everyday conversation, say, the different interpretations would not matter. We compared the sensitivity of the differences between the logical operators by the Kullback-Leibler distances between their distributions. The distance of an inferred distribution, inferred from a logical argument, from the

[6]In the literature Spearman correlation copulas are often preferred to Gaussian copulas as they keep the distribution of the marginals and the correlation independent.

uniform distributions, as a standard of ignorance, is an indicator of the informativeness and strength of the argument.

We remembered that neglecting base rates may be rational under natural sampling conditions. This property holds for beta distributions, their expected values and variances. We have demonstrated how typical tasks of deductive reasoning (Rips, 1994) can be cast into a probabilistic format including imprecision. A paradoxical property is observed in Doherty's information seeking task (Doherty et al., 1979; Tweney et al., 2010; Kleiter, 2015): Sampling more and more information from just one experimental condition, without sampling from a control condition, leads to less and less precise conclusions. The suppression task (Byrne, 1989) was among the first tasks framed and analyzed in a probabilistic format (Stevenson and Over, 1995). Expressing the implicit assumptions by second order probability distribution predicts the empirical results reported in the literature. Jeffrey's proposal of how to update probabilities by uncertain evidence is well known as Jeffrey's rule (Jeffrey, 1965). In a bivariate model with two first order probabilities X and Y treated as random variables the problem becomes a typical regression problem, predicting the distribution of Y given a value of X.

Gigerenzer et al. (1991) proposed a probabilistic mental model (PMM) of confidence judgments. The model was introduced and demonstrated by the experimental paradigm of *city size judgments*. In the first of two experiments twenty five German cities with more than 100,000 inhabitants were selected. Participants were presented all 300 pairs of the cities and asked to decide which one has more inhabitants. In addition, the participants rated how sure they were that each of their choices was correct.

Using just one quantitative property, city size, underlying all questions in the experimental procedure introduced a big difference with respect to the general knowledge almanac questions widely used in other studies of overconfidence[7].

The data may be looked at from the perspective of the method of paired comparison (Thurstone, 1927). Processing the data with Thurstone's probabilistic model of paired comparison one would introduce a normal distribution for the size of each of the cities. Such a probability distribution models the participant's knowledge about the size of a city and the precision of this knowledge. The confidence judgment then becomes a function of the differences in the location and spread of these distributions. The distributions are thus not second order probability distributions, but distributions over a quantitative property, here the number of inhabitants of a city. The property is imprecise (compare the intervals in Figure 2 of Gigerenzer et al., 1991), not the probability[8]. The same holds for the cues in the PMMs.

I consider the analyses presented in this contribution as part of a thorough task analysis of reasoning tasks. Task analysis is a prerequisite for a good psychological investigation. The results of our analyses show how difficult it may be to run a good reasoning experiment. A major problem, e.g., is how to manipulate and measure imprecision. Another problem is that inferences with the same logical operators or the same logical inference rules may be different for different levels of the probabilities of the premises. High probabilities may lead to one result, low probabilities to a different one. Results may also not be invariant with respect to positive or negative correlations of the involved uncertain quantities and risks.

Modeling imprecise judgments has a long history. It started with Gauss and his analysis of human judgment errors in astronomical observations. It continued in the nineteenth century with Weber's and Fechner's just noticeable differences, thresholds and psychophysical functions. The probabilistic modeling of sensory data by von Helmholtz pioneered present day's Free Energy Principle. Thorndike introduced the law of comparative judgment. In the second half of the twentieth century signal detection theory, stimulus sampling theory, stochastic choice theory, Brunswick's lens model, stochastic response models, neural networks, and decision theory took up the problem. At the beginning of the twenty first century computational neuroscience contributed substantially to model imprecision in information processing.

Models of the functioning of the brain claim that the neuronal processes underlying cognitive processes like memory, perception, or decision making are inherently *stochastic* and *noisy*. A good example is the work of Rolls and Deco (2010). Spike trains of neurons follow Poisson distributions, cell assemblies are modeled by mean-field analysis and the dynamics of elementary decision processes are simulated by integrate-and-fire neural networks. The authors observe that "... if a decision must be made based on one's confidence about a decision just made, a second decision-making network can read the information encoded in the firing rates of the first decision-making network to make a decision based on confidence ..." (Rolls and Deco, 2010, p. 167). A probability assessment is a read-out of one's own confidence, the product of an auto-epistemic self-monitoring process (Rolls and Deco, 2010, p.196ff.). The assessment might correspond to the point of maximum probability of being coherent.

Precision plays an important role in the theories of free energy, active inference, and predictive coding (Friston, 2010; Buckley et al., 2017). In a task in which the participants had to decide on the direction of a set of systematically moving dots in a set of randomly moving dots the precision of the responses was related to the response times. It was shown that the precision of the responses was controlled (among other locations) in the posterior parietal cortex (FitzGerald et al., 2015). Precision may be modulated by neurotransmitters. Friston et al. (2012), for example, hypothesized that precision is related to dopamin.

[7]The study of overconfidence can be tricky as overconfidence for E goes hand in hand with underconfidence for non-E. Scoring rules avoid this problem (Kleiter et al., 2002).

[8]It may be mentioned that the evaluation of the data by the method of paired comparison would allow to calculate several interesting statistics like item characteristics, the consistency of the judgments, or interindividual differences.

In probability logic all operators and inference rules infer interval probabilities. Using conclusions iteratively would require to propagate lower and upper probabilities again and again. For a human brain to keeping track of lower and upper bounds will soon become too messy. One way out of the exploding complexity is to simplify and process the probability distributions of being coherent. To use a metaphor: In a cell assembly the distributions may result from the many single cell activations constrained by the coherence criterion.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02051/full#supplementary-material

REFERENCES

Adams, E. W. (1965). The logic of conditionals. *Inquiry* 8, 166–197. doi: 10.1080/00201746508601430

Adams, E. W. (1966). "Probability and the logic of conditionals," in *Aspects of Inductive Logic*, eds J. Hintikka and P. Suppes (Amsterdam: North-Holland), 265–316.

Aitchison, J., and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge: Cambridge University Press.

Augustin, T., Colen, F. A., de Cooman, G., and Troffaes, M. C. M., editors (2014). *Introduction to Imprecise Probabilities*. Chichester: Wiley.

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

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

Bayes, T. R. (1763/1958). An essay towards solving a problem in the doctrine of chance. *Biometrika* 45, 293–315. doi: 10.1093/biomet/45.3-4.296

Bellantuono, I. (2018). Find drugs that delay many diseases of old age. *Nature* 554, 293–295. doi: 10.1038/d41586-018-01668-0

Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. (2017). The free energy principle for action and perception: a mathematical review. *J. Math. Psychol.* 81, 55–79. doi: 10.1016/j.jmp.2017.09.004

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition* 39, 61–83. doi: 10.1016/0010-0277(89)90018-8

Cuzzolin, F. (2012). "Generalizations of the relative belief transform," in *Belief Functions: Theory and Applications. 2nd International Converence on Belief Functions*, eds T. Denoeux and M. Masson (Berlin: Springer), 109–116.

Diaconis, P., and Zabell, S. L. (1982). Updating subjective probability. *J. Am. Stat. Assoc.* 77, 822–830. doi: 10.1080/01621459.1982.10477893

Doherty, M. E., Chadwick, R., Garavan, H., Barr, D., and Mynatt, C. R. (1996). On people's understanding of the diagnostic implications of probabilistic data. *Mem. Cogn.* 24, 644–654.

Doherty, M. E., Mynatt, C. R., Tweney, R. D., and Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychol.* 43, 111–121. doi: 10.1016/0001-6918(79)90017-9

Donkin, W. F. (1851). On certain questions relating to the theory of probabilities. *Philos. Mag.* 1, 353–368. doi: 10.1080/14786445108646751

Draheim, D. (2017). *Generalized Jeffrey Conditionalization. A Frequentist Semantics of Partial Conditionalization*. Cham: Springer.

Elqayam, S. (2017). "New psychology of reasoning," in *International Handbook of Thinking and Reasoning*, eds L. J. Ball and V. E. Thompson (Hove: Routledge), 130–150.

Erev, I., Wallsten, T. S., and Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychol. Rev.* 101, 519–527. doi: 10.1037/0033-295X.101.3.519

FitzGerald, T. H. B., Moran, R. J., Friston, K. J., and Dolan, R. J. (2015). Precision and neuronal dynamics in the human posterior parietal cortex during evidence accumulation. *NeuroImage* 107, 219–228. doi: 10.1016/j.neuroimage.2014.12.015

Friedman, J. A., Baker, J., Mellers, B. A., Tedlock, P. E., and Zeckhauser, R. (2018). The value of precision in probability assessment: evidence from a large-scale geopolitical forecasting tournament. *Int. Stud. Q.* 62, 410–422. doi: 10.1093/isq/sqx078

Friston, K. (2010). The free-energy principle: a rough guide to the brain? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., et al. (2012). Dopamin, affordance and active inference. *PLoS Comput. Biol.* 8:e1002327. doi: 10.1371/journal.pcbi.1002327

Fugard, A. J., Pfeifer, N., Mayerhofer, B., and Kleiter, G. D. (2011). How people interpret conditionals: shifts toward the conditional event. *J. Exp. Psychol.* 37, 635–648. doi: 10.1037/a0022329

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 Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychol. Rev.* 98, 509–528. doi: 10.1037/0033-295X.98.4.506

Gilio, A. (1995). "Algorithms for precise and imprecise conditional probability assessments," in *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence*, eds G. Coletti, D. Dubois, and R. Scozzafava (Planum Press, New York), 231–254.

Gilio, A. (2012). Generalization of inference rules in coherence-based probabilistic default reasoning. *Int. J. Approx. Reason.* 53, 413–434. doi: 10.1016/j.ijar.2011.08.004

Gilio, A., and Sanfilippo, G. (2013). "Conditional random quantities and iterated conditioning in the setting of coherence," in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, ed L. Van der Gaag (Berlin: Springer), 218–229.

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

Gupta, A. K., and Nadarajah, S., editors (2004). *Handbook of Beta Distribution and Its Application*. New York, NY: Marcel Dekker.

Hadjichristidis, C., Sloman, S. A., and Over, D. E. (2014). *Categorical Induction From Uncertain Premises: Jeffrey's Doesn't Completely Rule*. Technical report, Department of Economics and Management, University of Trento (Trento).

Halpern, J. Y. (2003). *Reasoning About Uncertainty*. Cambridge, MA: MIT Press.

Jaynes, E. T. (2003). *Probability Theory. The Logic of Science*. Cambridge: Cambridge University Press.

Jeffrey, R. (1965). *The Logic of Decision*. New York, NY: McGraw-Hill.

Jeffrey, R. (1992). *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.

Jeffrey, R. (2004). *Subjective Probability. The Real Thing*. Cambridge: Cambridge University Press.

Johnson, N. L., and Kotz, S. (1970). *Continuous Univariate Disbrigugions*, Vol. 2. Boston, MA: Hoghton Mifflin.

Johnson-Laird, P. N., and Shafir, E. (1993). The interaction between reasoning and decision making: an introduction. *Cognition* 49, 1–9. doi: 10.1016/0010-0277(93)90033-R

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

Kleiter, G. D. (1981). *Bayes-Statistik. Grundlagen und Anwendungen*. Berlin: Walter de Gruyter.

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.

Kleiter, G. D. (1996). Critical and natural sensitivity to base rates [comments to Koehler (1996)]. *Behav. Brain Sci.* 19, 27–29. doi: 10.1017/S0140525X00041297

Kleiter, G. D. (2013). "Ockham's razor in probability logic," in *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, Advances in Intelligent Systems and Computation, 190, eds R. Kruse, M. R. Berthold, C. Moewes, M. A. Gil, P. Grzegorzewski, and O. Hryniewicz (Heidelberg: Springer), 409–417.

Kleiter, G. D. (2015). Modeling biased information seeking with second order probability distributions. *Kybernetika* 51, 469–485. doi: 10.14736/kyb-2015-3-0469

Kleiter, G. D. (2018). Adams' p-validity in the research on human reasoning. *J. Appl. Logics* 5, 775–825.

Kleiter, G. D., Doherty, M. E., and Brake, G. L. (2002). "The psychophysics metaphor in calibration research," in *Frequency Processing and Cognition*, eds P. Sedlmeier and T. Betsch (Oxford: Oxford University Press), 239–255.

Kleiter, G. D., Fugard, A. J. B., and Pfeifer, N. (2018). A process model of the understanding of uncertain conditionals. *Think. Reason.* 24, 386–422. doi: 10.1080/13546783.2017.1422542

Koehler, J. J. (1996). The base rate fallacy reconsidered: descriptive, normative, and methodological challenges. *Behav. Brain Sci.* 19, 1–53. doi: 10.1017/S0140525X00041157

Kurowicka, D., and Cooke, R. (2004). "Distribution-free continuous Bayesian belief nets," in *Proceedings of the Fourth International Conference on Mathematical Methods in Reliability Methodology and Practice* (Santa Fe).

Kurowicka, D., and Cooke, R. (2006). *Uncertainty Analysis With High Dimension Dependence Modelling*. Chichester: Wiley.

Kurowicka, D., and Joe, R. (2011). *Dependence Modeling: Vine Copula Handbook*. Singapure: World Scientific.

Lad, F. (1996). *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction*. New York, NY: Wiley.

Levi, I. (1967). Probability kinematics. *Brit. J. Philos. Sci.* 18, 197–209. doi: 10.1093/bjps/18.3.197

Lichtenstein, S., and Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychon. Sci.* 9, 563–564. doi: 10.3758/BF03327890

Mai, J.-F., and Scherer, M. (2012). *Simulating Copulas. Stochastic Models, Sampling Algorithms, and Applications*. London: Imperial College Press.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Berlin: Springer.

Oaksford, M., and Chater, N. (1995). Information gain explains relevance which explains the selection task. *Cognition* 57, 97–108. doi: 10.1016/0010-0277(95)00671-K

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

Over, D. (2009). New paradigm psychology of reasoning. *Think. Reason.* 15, 431–438. doi: 10.1080/13546780903266188

Peterson, C. R., and Beach, L. R. (1967). Man as an intuitive statistician. *Psychol. Bull.* 68, 29–46. doi: 10.1037/h0024722

Pfeifer, N., and Kleiter, G. D. (2005). Towards a mental probability logic. *Psychol. Bel.* 45, 71–99. doi: 10.5334/pb-45-1-71

Pfeifer, N., and Kleiter, G. D. (2006a). "Towards a probability logic based on statistical reasoning," in *Proceedings of the 11th IPMU Conference* (Paris), 2308–2315.

Pfeifer, N., and Kleiter, G. D. (2006b). Inference in conditional probability logic. *Kybernetika* 42, 391–404.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rips, L. J. (1994). *The Psychology of Proof. Deductive Reasoning in Human Thinking*. Bradford; Cambridge, MA: MIT Press.

Rolls, E. T., and Deco, G. (2010). *The Noisy Brain. Stochastic Dynamics as a Principle of Brain Function*. Oxford: Oxford University Press.

Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Nagler, T., Erhardt, T., et al. (2018). *Statistical Inference of Vine Copulas*. Software.

Smets, P. (1990). "Constructing the pignistic probability function in a context of uncertainty," in *Uncertainty in Artificial Intelligence*, Vol. 5 (Amsterdam: North Holland), 29–40.

Smets, P., and Kruse, R. (1997). "Imperfect information: imprecision and uncertainty," in *Uncertaint Management in Information Systems*, eds A. Motro and P. Smets (Boston, MA: Kluwer), 343–368.

Spiegelhalter, D. J., Franklin, R. C. G., and Bull, K. (1990). "Assessment, criticism and improvement of imprecise subjective probabilities for a medical expert system," in *Uncertainty in Artificial Intelligence 5*, eds M. Henrion, R. D. Shachter, L. N. Kanal, and J. F. Lemmer (Amsterdam: North-Holland), 285–294.

Sriboonchitta, S., Wong, W.-K., Dhompongsa, S., and Nguyen, H. R. (2010). *Stochastic Dominance and Applications to Finance, Risk and Economics*. Boca Raton, FL: CRC, Taylor & Francis.

Staël von Holstein, C.-A., S. (1970). *Assessment and Evaluation of Subjective Probabililty Distributions*. Stockholm: The Economic Research Institute at the Stockholm School of Economics.

Stevenson, R. J., and Over, D. E. (1995). Deduction from uncertain premises. *Q. J. Exp. Psychol.* 48, 613–643. doi: 10.1080/14640749508401408

Suppes, P. (1966). "Probabilistic inference and the concept of total evidence," in *Aspects of Inductive Logic*, eds J. Hintikka and P. Suppes (Amsterdam: North-Holland), 49–65.

Tenenbaum, J. B., Griffiths, T. L., and Niyogi, S. (2007). "Intuitive theories as grammars for causal inference," in *Causal Learning: Psychology, Philosophy, and Computation*, eds A. Gupnik, and L. Schulz (Oxford: Oxford University Press), 301–322.

Thurstone, L. L. (1927). The method of paired comparisons for social values, 21, (1927), 384-400. *J. Abnorm. Soc. Psychol.* 21, 384–400. doi: 10.1037/h0065439

Tversky, A., and Kahneman, D. (1983). Extension versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90, 293–315. doi: 10.1037/0033-295X.90.4.293

Tweney, R. D., Doherty, M. E., and Kleiter, G. D. (2010). The pseudodiagnosticity trap. Should subjects consider alternative hypotheses? *Think. Reason.* 16, 332–345. doi: 10.1080/13546783.2010.525860

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Wallmann, C., and Kleiter, G. (2014a). Probability propagation in generalized inference forms. *Studia Logica* 102, 913–929. doi: 10.1007/s11225-013-9513-4

Wallmann, C., and Kleiter, G. D. (2012a). "Beware of too much information," in *Proceedings of the 9th Workshop on Uncertainty Processing, WUPES*, eds T. Kroupa and J. Vejnarova (Prague: Fakulty of Managment, University of Economics), 214–225.

Wallmann, C., and Kleiter, G. D. (2012b). "Exchangeability in probability logic," in *Communications in Computer and Information Science, IPMU (4)*, Vol. 300, eds S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo, and R. R. Yager (Berlin: Springer), 157–167.

Wallmann, C., and Kleiter, G. D. (2014b). Degradation in probability logic: when more information leads to less precise conclusions. *Kybernetika* 50, 268–283. doi: 10.14736/kyb-2014-2-0268

Wedlin, A. (1996). "Some remarks on the transition from a standard Bayesian to a subjectivistic statistical standpoint," in *Proceedings of the "International Conference, The Notion of Event in Probabilistic Epistemology". Applicata "Bruno de Finetti"* (Triest: Dipartimento di Matematica), 91–110.

Wuebbles, D. J., Fahey, D. W., Hibbard, K. A., Dokken, B. C., Stewart, B. C., and Maycock, T. K. (eds.). (2017). *Climate Science Special Report: Fourth National Climate Assessment*, Vol. 1. Washington, DC: U. S. Global Change Research Program.

Xu, F., and Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychol. Rev.* 114, 245–272. doi: 10.1037/0033-295X.114.2.245

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: 27 March 2018
doi: 10.3389/fpsyg.2018.00403

The Role of Type and Source of Uncertainty on the Processing of Climate Models Projections

Daniel M. Benjamin[1]* *and David V. Budescu*[2]

[1] *Biomedical Ethics Unit, Department of Social Studies of Medicine, McGill University, Montreal, QC, Canada,*
[2] *Department of Psychology, Fordham University, New York, NY, United States*

Scientists agree that the climate is changing due to human activities, but there is less agreement about the specific consequences and their timeline. Disagreement among climate projections is attributable to the complexity of climate models that differ in their structure, parameters, initial conditions, etc. We examine how different sources of uncertainty affect people's interpretation of, and reaction to, information about climate change by presenting participants forecasts from multiple experts. Participants viewed three types of sets of sea-level rise projections: (1) precise, but *conflicting*; (2) *imprecise*, but agreeing, and (3) *hybrid* that were both conflicting and imprecise. They estimated the most likely sea-level rise, provided a range of possible values and rated the sets on several features – ambiguity, credibility, completeness, etc. In Study 1, everyone saw the same hybrid set. We found that participants were sensitive to uncertainty between sources, but not to uncertainty about which model was used. The impacts of conflict and imprecision were *combined for estimation tasks* and *compromised for feature ratings*. Estimates were closer to the experts' original projections, and sets were rated more favorably under imprecision. Estimates were least consistent with (narrower than) the experts in the hybrid condition, but participants rated the conflicting set least favorably. In Study 2, we investigated the hybrid case in more detail by creating several distinct interval sets that combine conflict and imprecision. Two factors drive perceptual differences: *overlap* – the structure of the forecast set (whether intersecting, nested, tangent, or disjoint) – and a*symmetry* – the balance of the set. Estimates were primarily driven by asymmetry, and preferences were primarily driven by overlap. Asymmetric sets were least consistent with the experts: estimated ranges were narrower, and estimates of the most likely value were shifted further below the set mean. Intersecting and nested sets were rated similarly to imprecision, and ratings of disjoint and tangent sets were rated like conflict. Our goal was to determine which underlying factors of information sets drive perceptions of uncertainty in consistent, predictable ways. The two studies lead us to conclude that perceptions of agreement require intersection and balance, and overly precise forecasts lead to greater perceptions of disagreement and a greater likelihood of the public discrediting and misinterpreting information.

Keywords: sources of uncertainty, conflict, imprecision, climate change, global warming, forecasting, ambiguity, vagueness

OPEN ACCESS

Edited by:
Nathan Dieckmann,
Oregon Health & Science University, United States

Reviewed by:
Mirta Galesic,
Santa Fe Institute, United States
Lin Guo,
University of Pennsylvania, United States

***Correspondence:**
Daniel M. Benjamin
daniel.benjamin2@mail.mcgill.ca

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

Received: *04 November 2017*
Accepted: *12 March 2018*
Published: *27 March 2018*

Citation:
Benjamin DM and Budescu DV (2018) The Role of Type and Source of Uncertainty on the Processing of Climate Models Projections. Front. Psychol. 9:403. doi: 10.3389/fpsyg.2018.00403

INTRODUCTION

Climate forecasts are riddled with uncertainty because climate models involve uncertainties around the model's structure, the measurement of initial conditions, the parameters of the key variables (e.g., future radiative forcing, population growth, economic activity), and the relationship between these variables. Moreover, because of the interactions between these uncertainties, models are typically run multiple times with different initial conditions and parameterizations, generating a spectrum of predictions to properly capture the deep uncertainties that drive the phenomena. The communication of such deep uncertainty is crucial to allow decision-makers (DMs) to make choices based on an accurate understanding of the state-of-the-art science and strength of the evidence (e.g., Drouet et al., 2015). If scientists do not properly communicate the nature, sources, and magnitude of the uncertainties, the DMs can be either over- or under-confident in the evidence and, in many cases, this can lead to suboptimal decisions (Fischhoff and Davis, 2014). The effects of poorly specified uncertainty can be profound. For example, the North Carolina Sea-Level Rise Assessment Report (N.C. Coastal Resources Commission's Science Panel on Coastal Hazards, 2010) projected a 39-inch rise in sea-level (ranging from 15 to 55 inches) in the Outer Banks by 2100. In response to this overly precise, long term projection, local conservative groups, worried about the economic devastation associated with this projection, launched an effective campaign against policy initiatives. The local government subsequently banned policy addressing these sea level projections suggesting much valuable real estate would be under water (Siceloff, 2014, News and Observer).

Climate Model Complexity and Decision-Making

Although there is high agreement among experts about the reality and causes of climate change (CC) (e.g., Doran and Zimmerman, 2009), there is much less agreement among projections of the future climate. Experts disagree on the primary drivers of uncertainty in climate projections including if and how such vital uncertainties can be resolved (e.g., Morgan and Keith, 1995; Zickfeld et al., 2010). For model projections to be useful, stakeholders in areas as diverse as biodiversity, water, transportation, energy, and city and regional planning must resolve the indeterminacy stemming from multiple experts running multiple models with multiple initial conditions producing multiple projections.

When making decisions under deep uncertainty, Decision Makers (DMs) form mental models of the complex systems involved, and these mental models drive subsequent beliefs and behaviors (Newell and Pitman, 2010; Galesic et al., 2016). When mental models are established, even tentatively, DMs evaluate and fit new information into their existing structure and beliefs. Holyoak and Simon (1999) have empirically demonstrated this process for legal decisions: Once an individual reaches a tentative decision, subsequent evaluations of evidence and arguments are affected by the original decision, which in turn influences future decisions. People also distort information to fit their tentatively favored alternative (e.g., Russo and Yong, 2011). Ambiguity in the definition of events, as well as vagueness and imprecision in projected outcomes, allows DMs to interpret results congruently with their own mental models instead of altering their beliefs to incorporate the full range of information (see Kunda, 1990).

Sources of Uncertainty

The problem of subjective interpretation is magnified when information comes from multiple sources. Research distinguishes between two sources of indeterminacy stemming from multiple sources: conflict and imprecision (Smithson, 1999, 2015). *Imprecision* (sometimes referred to as ambiguity or vagueness) occurs when quantities are specified inexactly and often takes the form of a range of possible outcomes (e.g., "We expect 1–3 inches of snow in the next 24 h") or an approximation ("We expect about 2 inches of snow"). *Conflict* occurs when quantities cannot simultaneously hold true ("Expert A expects 1 inch of snow in the next 24 h" and "Expert B expects 3 inches"). DMs are generally more conflict averse than imprecision averse (Smithson, 1999), but both conflict and imprecision contribute toward overall perceptions [operationalized by subjective ratings of uncertainty (Smithson, 2015)].

Professionals, such as insurance underwriters and actuaries instinctively differentiate between these sources of indeterminacy. Cabantous et al. (2011) presented insurers with risk estimates for three hazardous events –fires, floods, and hurricanes – from two modeling firms. The models agreed on a mean value (risk), disagreed at either end of a range of values (conflict), or agreed over the same range of values (imprecision). The insurers tended to charge higher premiums for catastrophic risks (e.g., floods) under conflict and higher premiums for non-catastrophic risk (e.g., house fires) under imprecision.

Although imprecision and conflict can operate simultaneously, previous research has focused on the extreme cases where they are distinct. The current studies examine how various sources of uncertainty impact how DMs aggregate, process, and resolve uncertain information from multiple sources in the context of projections related to CC.

Attitudes Toward Imprecision

Decision-makers generally prefer precise over imprecise options (Ellsberg, 1961; Wallsten et al., 1993; Kramer and Budescu, 2004), but are sensitive to the level of precision and resolution that can be expected in different contexts. As a rule, DMs prefer the most precise option that can be reasonably expected within a specific context. The congruence principle (Wallsten and Budescu, 1995) states that DMs seek congruence between the degree of precision of an event, the nature of the uncertainty surrounding the event, and the representation of the uncertainty. For example, DMs expect very precise estimates of uncertainty for unambiguous events with easily quantifiable uncertainties (e.g., the chance that a man born and residing in the United States will live at least X years). On the other hand, they would probably reject equally precise estimates in the context of ambiguous events with hard to model and quantify uncertainties (e.g., the chance of a *substantial drop* in the national unemployment rate *in the foreseeable future*)

and would consider a moderately imprecise estimate to be more credible and informative.

There is also evidence that laypeople do not always prefer precision. Many individuals are imprecision (ambiguity) seeking for unlikely gains and for likely losses (Einhorn and Hogarth, 1985, 1988) for both outcomes and probabilities (Hogarth and Einhorn, 1990; Casey and Scholz, 1991; González-Vallejo et al., 1996; Budescu et al., 2002). For example, a preference for imprecision was demonstrated in financial forecasting by showing that DMs find moderately imprecise financial forecasts to be more credible, more accurate, and induce more confidence than their precise counterparts (Du and Budescu, 2005; Du et al., 2011). Because of the complexity of predicting the future climate, DMs would not expect highly precise predictions (such as a point estimate of the mean global temperature over 50 years). In a task inspired by CC, DMs who could use one of two decision aids, preferred one that graphically showed the full range of values (i.e., stressing and highlighting uncertainty) over one that calculated the expected value of the options and eliminated all uncertainty (Budescu et al., 2014a,b).

Vagueness as Conflict

Vagueness in complex domains is often driven by expert disagreement. Experts fail to arrive at the same conclusion (whether precise or not) when there are too many unresolved or unknown relationships among variables. The belief that scientists disagree about severity and causes of climate change decreases the endorsement of corrective actions, including policy initiatives, to address the problem. Lewandowsky et al. (2013) show that explaining that scientists agree that humans are causing climate change, increases agreement that climate change and certain climate trends (increased temperature, sea-level, and natural disasters) are attributable to human activity. Differences in perceptions of a scientific consensus are driven by individuals' worldview, measured both by the strength of their belief in the free-market (Lewandowsky et al., 2013) and their cultural cognition – a theory describing how the values associated with cultural identity determine beliefs (Kahan et al., 2011).

Decision-makers must decide how to weight competing experts' forecasts based on their information, knowledge, ability, beliefs, etc. Disagreement among experts can be attributed to features of the experts, such as competence, knowledge, bias, or their candidness about uncertainty and to environmental factors, such as complexity and stochasticity (Hammond, 1996; Shanteau, 2000; Dieckmann et al., 2017). Interestingly, the public's knowledge and ability drive their perceptions of the experts' knowledge and ability, so that DMs with less topic knowledge and who are less numerate are more likely to attribute expert disagreement to incompetence (Dieckmann et al., 2015). More knowledgeable DMs attribute the conflict to various biases and conflicts of interest, while more numerate DMs attribute it to the stochastic nature of the events. DMs must reconcile disagreeing forecasts by aggregating the available information with their own beliefs. When individual judges combine forecasts, they are sensitive to the structure of the information and the nature of their cognitive processes (Wallsten et al., 1997). Simple aggregation methods, like averaging, are often highly accurate and robust (Clemen, 1989). However, DMs often fail to understand the benefits of averaging for reducing individual error (Larrick and Soll, 2006), and fail to adjust their own beliefs sufficiently to incorporate the advice of others (Yaniv and Milyavsky, 2007).

Communication of Vague Information

The presentation of uncertain information is a tradeoff between providing enough precision to be useful while being sufficiently imprecise to be realistic. The communicator and audience often have competing goals. Communicators prefer to communicate vaguely, and audiences prefer precise information (e.g., Erev and Cohen, 1990; Wallsten et al., 1993). The description of uncertainty around climate change has ultimately led to a divide between public and scientific perceptions of the problem (Zehr, 2016), even though greater uncertainty leads to a greater expectation of risk and damage (Lewandowsky et al., 2014).

Risk communication experts recommend transparency about uncertainty to aid interpretability for DMs (Fischhoff and Davis, 2014). A simple and common communication tool is to provide range estimates, such as confidence intervals, to express the scope of values considered reasonably possible or probable. Uncertainty about climate change, when presented as a range estimate, is considered more credible when certainty is not possible (MacInnis et al., unpublished). Interval estimates are perceived to be more credible in hindsight and to have higher utility for deciding at higher likelihoods (Dieckmann et al., 2010). Vagueness that characterizes both numerical ranges and verbal ambiguity, interacts with message framing resulting in vagueness avoidance for positively framed values and vagueness seeking for negatively framed values (Kuhn, 1997). Ranges can improve attributions of the likelihoods across possible outcomes (Dieckmann et al., 2012) and can improve the appropriateness of steps taken to address weather-related risks (Joslyn and LeClerc, 2012).

The Current Paper

We examine DMs' reactions to sources of uncertainty arising from multiple forecasts in the context of CC. Climate is a perfect domain for such studies. Due to the computational constraints of running complex climate models, it is often impossible to resolve these disagreements and indeterminacies during modeling. There is a tradeoff between model resolution and expected accuracy in climate models. Modelers can set model parameters to estimate the future climate with high resolution (say at the level of a county or a neighborhood), but higher resolution can reduce confidence in the accuracy of their forecasts. Conversely, lower resolution forecasts (e.g., global models) may be perceived as more accurate but uninformative or even irrelevant at a local level. This tradeoff is understood by laypeople who intuitively rate narrower intervals as more informative than their precise counterparts (e.g., Du et al., 2011), but less likely to be accurate, and vice versa (Yaniv and Foster, 1995). Therefore, to maintain a high degree of confidence, climate projections are typically expressed with equivocacy by including vagueness or uncertainty information.

In two studies, we present respondents with two forecasts related to future impacts of CC with various sources of indeterminacy and disagreement. Consistent with the CC literature we differentiate between *model and source uncertainty*. Model uncertainty describes indeterminacy stemming from the models' structure and inputs, and source uncertainty describes disagreement between how expert sources utilize models and interpret their output. The forecasts – which are projections of sea level rise and their impact on the ports in southern California over the next 50 years – vary in whether they came from the same or different models and use the same or different initial conditions. In addition, the various sets of experts' forecasts reflect different degrees of imprecision and conflict. In each case, the respondents estimated the most likely value as well as a range of possible values. Additionally, we obtained their confidence in those estimates in two ways: direct ratings and willingness-to-pay for insurance to reduce the risk of losing money when betting on their estimates. Finally, they provided comparative ratings of the forecast sets on key attributes.

STUDY 1

We test how DMs react to various sources and types of uncertainty underlying model projections. More specifically, we test a 3 × 4 typology of uncertainty. One factor consists of three levels of *source uncertainty* that describe how forecasts from two experts relate to each other – conflict between precise experts, imprecise agreeing experts, and a hybrid case which includes both conflict and imprecision. The second factor – *model uncertainty* – is based on a 2 × 2 crossing of *structural uncertainty* about the model and uncertainty reflecting *judgmental* interpretation of the models' results. We manipulate structural uncertainty by providing projections from the same, or different, models, and we manipulate judgmental uncertainty by providing projections from one or two sets of initial conditions. See **Table 1** for a schematic description of the design.

We test the effects of these various facets of uncertainty on three distinct sets of dependent variables. The first two sets – estimates of the target quantity and its feasible bounds, and confidence in the estimates – are obtained for each case. The third set consists of a group of concurrent, comparative ratings of the various cases obtained after making all estimates. This set provides a retrospective global evaluation of the various sources of uncertainty.

Overall, our goal is to test how sensitive the DMs' estimates and expressions of confidence will be to the manipulations of the various facets of uncertainty. Consider first the effects of the source of uncertainty. We do not expect any differences in terms of the best estimates of the target quantity which will, invariably, be some simple aggregate of, or compromise between, the two extreme – the lowest and the highest – values presented by the forecasters[1].

We have differential expectations about the range of the estimates and the confidence they inspire. We expect that the DMs will be most faithful to (deviate least from) the experts' estimates in the case of identical and imprecise forecasts, and we expect that the range of estimates will be the narrowest in the hybrid case that, by its nature, highlights a narrow area of agreement, so we expect the DMs to focus on it. In line with previous results (Smithson, 1999; Baillon et al., 2012), we expect the lowest levels of confidence and the highest willingness to purchase insurance in the presence of conflict between the experts. Similarly, we expect that, in retrospect, DMs will rate the conflicting cases lowest (least desirable or attractive) in all respects, and rate the imprecise cases highest on most attributes, with the hybrid cases in between.

We expect that DMs will be less sensitive to our manipulations of model uncertainty which are more subtle and, unlike source uncertainty which is very direct and salient, reflect a deeper understanding and analysis of the situation. In general, we expect that lower (structural and judgmental) uncertainty will induce narrower ranges of estimates, higher levels of confidence and less willingness to purchase insurance. Finally, we predict that DMs will react more strongly to information that reduces *structural* (rather than *judgmental*) uncertainty, because structural uncertainty implies there is conflict in the relationship between physical phenomena while judgmental uncertainty, can be attributed to changes in initial conditions or model parameters, and can be resolved more easily.

Methods

Participants and Design

One hundred and thirty undergraduate and graduate students from New York University and Fordham University participated in the study. Eighty-seven (67%) were female, and the mean age was 20.9 (SD = 2.8). Participants received $10 for participating and a performance-based prize (up to an additional $20;

[1]Since there are only two values, X_{min} and X_{max}, their mean, median and mid-range are identical.

TABLE 1 | Typology of source and model uncertainty.

Model uncertainty (Between subjects)		Source of uncertainty (Randomly ordered within subjects)		
Structural uncertainty	**Judgmental uncertainty**	***Conflict* between two precise forecasts**	**Two identical *Imprecise* forecasts**	***Hybrid*: Two conflicting imprecise forecasts**
One model	One set of initial conditions			
	Two sets of initial conditions			
Two Models	One set of initial conditions			
	Two sets of initial conditions			

described below). We varied model uncertainty – a combination of two binary factors reflecting structural and judgmental uncertainties – between subjects. Participants were randomly assigned to one of four groups, and they saw projections from either one or two model(s), reflecting structural uncertainty, and using one or two set(s) of initial conditions, capturing judgmental uncertainty. We varied source uncertainty within-subjects, so all participants saw three projection sets from various pairs of experts: one set of precise, but conflicting projections; one set of imprecise, but agreeing, projections; and one set of *hybrid* projections that were both conflicting and imprecise (see **Table 1**).

Material and Procedure

Participants saw the same scenario – describing the effects of sea level rise on ports in Southern California – each time with a different set of projections (see Supplementary Materials for the full scenario). Additional tasks interspersed between them served as distractors from the manipulation. Each projection set consisted of two forecasts from two different experts. In each condition, the pairs of forecasts were attributed to the same two experts. These experts were labeled generically, so no specific valence or ideology could be ascribed to either expert. The three source uncertainty conditions were presented in random order; all three versions had the same mean (25 inches) and the same range (10–40 inches) of sea level rise projected over 50 years. The conflict condition consisted of disagreeing point estimates (10 vs. 40 inches), the imprecision condition consisted of agreeing interval estimates (both 10–40 inches), and the hybrid condition consisted of two non-overlapping intervals with a common endpoint (10–25 inches and 25–40 inches). To highlight why similar models could result in different output, the following statement was added to each scenario, "Note: Differences between projections may reflect the experts' uncertainty about the values of the key parameters."

Figure 1 presents the flow of the study from beginning to end. Participants started by reading background information about climate models including the basic science behind CC its potential impacts, a description of Earth System Models (EaSMs), and an explanation of why they are uncertain (see Supplementary Materials for introductory text). We developed this text in consultation with climate scientists, ecologists, and water experts. This information provided a basic understanding of the sources of imprecision and disagreement among the experts and model projections, even when they agree on the general science of CC.

Next, participants completed a 12-item belief in CC inventory. All items were five-option Likert statements (labeled from "strongly disagree" to "strongly agree") that were adapted from Heath and Gifford (2006). There were six subscales of perceptions of CC including general belief in, personal experience with, belief in humans causing, belief in serious consequences of, self-efficacy to make a difference in, and intentions to take alleviative actions to address, climate change. The original subscales include 4 items each, but we administered only 2 of the items to shorten the experiment time. We conducted a factor analysis on 374 responses (from the US sample of Budescu et al., 2014b) using a single factor and retained the top two items (i.e., with the highest loading) in each. The reduced scale maintained the level of reliability of the full scale and maintained or improved upon the reliability for all subscales except one (personal experience dropped from $\alpha = 0.87$ to 0.80; see Supplementary Table S1 in the Supplementary Material for the reduced scales). Participants

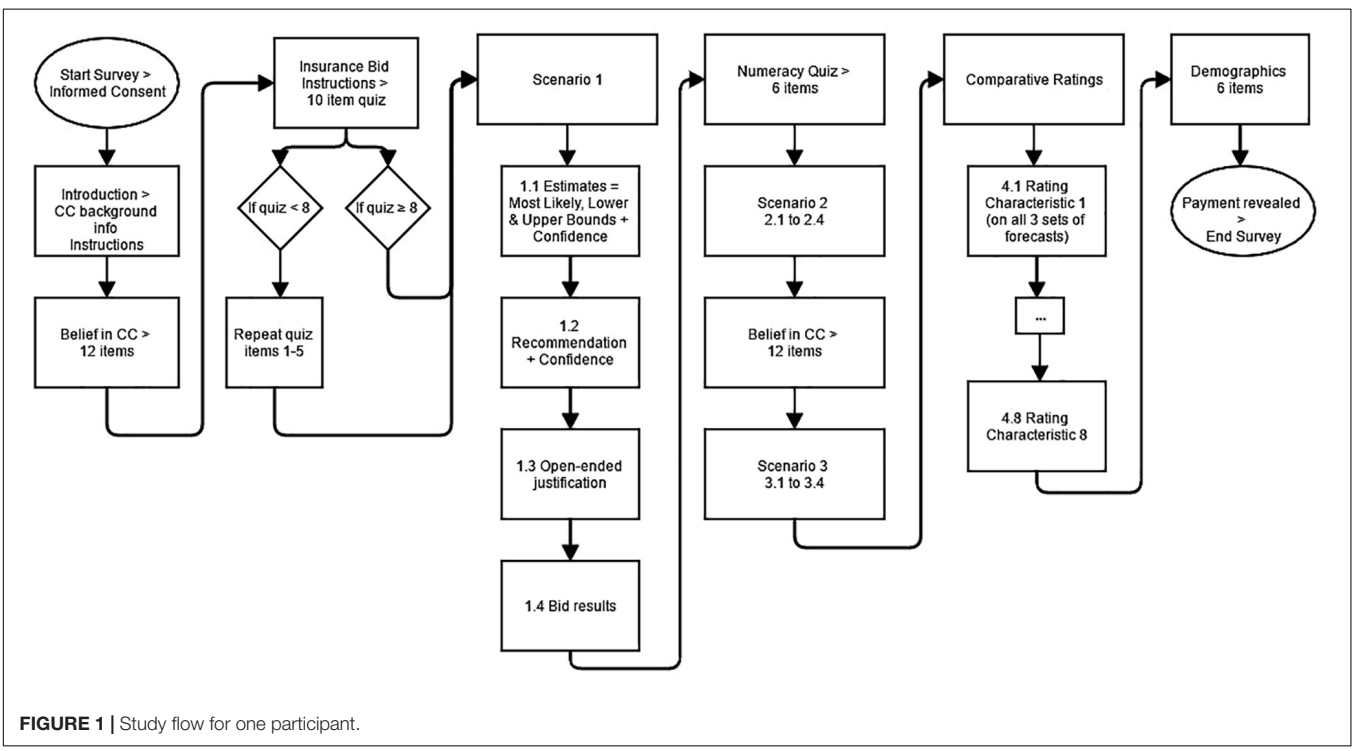


FIGURE 1 | Study flow for one participant.

were credited with an endowment of $10 for completing the 12 belief items for use in the first (of three) incentivized betting task to determine their underlying uncertainty.

Participants then read a one paragraph summary of a scenario regarding the need to raise the ports in Southern California to protect against projected sea level rise. Participants saw two expert projections pertaining to the scenario. The experts were labeled generically (e.g., Scientist A and B) and the models were given fictitious names Global Circulation Simulation model - version X (GCSX) and Earth System Generation model - version Y (ESGY).

After reading the scenario, the participants provided their best guess of the target value, stated a range of likely values, and rated their confidence on a seven-option Likert scale from "not at all" to "extremely." They then were asked to imagine they were a consultant to the port authority and recommend a value to plan for and rate their confidence on the same seven-point scale. They were also allowed to give an open-ended justification for their estimates, but an analysis of the content is beyond the scope of this paper.

To motivate the subjects to provide honest estimates, they were told that their estimates for a scenario will be part of a bet that potentially paid based on its accuracy. If their estimate was within 10% of the "true" value, they could double their $10 endowment, and if their estimate was off by more than 10%, they would lose their endowment. The true values were generated by calculating the mean of ten runs assuming expert projections was distributed normally over the range. Before resolving their bet, participants were offered the opportunity to purchase insurance in a bidding procedure adapted from Becker et al. (1964), by using the $10 they earned previously. Specifically, they could bid any amount $B ($0 $\leq$ B $\leq$ $7.50) to purchase insurance. If their bid was at least as large as our randomly generated counteroffer (between $0 and $5, in increments of $.25), they were insured, and their loss would be reduced to $2.50 plus the cost of insurance (which was equal to the counteroffer). If their bid was less than the counteroffer, they were uninsured. This procedure was designed to elicit bids that accurately reflect the DM's perceived uncertainty.

Participants were told that overbidding can lead to overspending on insurance and underbidding can lead to under-protection, and they completed a quiz about the bidding procedure. The quiz provided an example including a best guess, insurance bid, counteroffer, and true value. Participants were asked if the bid was successful in purchasing insurance, the price of insurance, if the bet was successful, winnings/losings, and total payment for the example. Participants answered these five questions for two different examples. If they answered fewer than 8 questions correctly, they repeated the first example and the associated quiz questions.

After completing all the stages of the first scenario, the participants took a six-item numeracy quiz which served as a distractor and to control for numeracy since it has been found to be a strong predictor of decision-making skill (e.g., Cokely et al., 2018). We adapted our numeracy quiz from the eight-item Weller et al. (2013) scale dropping two items, for being too time consuming. Participants were credited with a $10 endowment to use in the second bet for completing the numeracy quiz.

After completing all stages – estimates and insurance bids – of the second scenario, participants completed the same 12-item belief in CC inventory to test the reliability of this measure and as a distractor before the third and final projection set. They were credited with their third (and final) $10 endowment after the completion of this inventory and completed all stages – estimates and insurance bids – of the third scenario.

The estimation and insurance bid on the second and third procedure were identical to the original one. The labels identifying the experts and models and number of initial conditions were fixed across scenarios within participants. Only the values and structure of the forecasts varied across scenarios.

After completing the third projection set, participants concurrently rated the three projection sets on eight attributes: ambiguous, conflicting, precise, credible, likely to be accurate, informative, complete, and easy to reconcile/decide. All three projection sets were shown on the screen and subjects were asked to rate all three projection sets independently for each attribute using a 7-point Likert scale from "not at all" to "extremely." Each trait was presented on a separate screen in a random order.

Finally, the participants answered some basic demographic questions including age, sex, major of study, year in school, political affiliation (Republican, Independent, Democrat, or other), and strength of political identity (five-options from "very weak" to "very strong"), and they received the winnings of one randomly selected bet.

Results

We ran a series of 3 $\times$ 2 $\times$ 2 mixed MANCOVAs with source uncertainty as the within-subjects factor and (number of) models and initial conditions as between-subjects factors and numeracy as a covariate[2]. Several results stand out: (1) We did not find any significant differences across conditions for the participants' best estimates and their recommendations, as expected, so we report their personal estimates throughout the results; (2) We did not find significant effects of the two components of model uncertainty (structural and judgmental) on any dependent variable; (3) We found several systematic effects of source uncertainty which we describe and discuss below one dependent variable at a time (**Table 2** shows the means and standard errors of all estimates by source uncertainty); (4) We did not find order effects: response patterns do not change if we only analyze the first condition seen by each participant, so differences cannot be attributed to the influence of previous judgments.

Range Estimates Across Sources of Uncertainty

The lower bound, upper bound, and range estimates all varied significantly by source uncertainty. In all the conditions, the two estimated bounds were shifted away from the expert projections toward the center of the intervals. In other words, the judged lower bounds are higher than the lower forecasts,

[2]Interestingly, we did not find significant association between the DMs' responses with different levels of belief in climate change or political identification. Neither these factors, nor personal demographics were significant when included as covariates in the models.

the judged upper bounds are lower than the upper forecasts and, consequently, the judged ranges are narrower than the actual range of forecasts. The estimates vary significantly across conditions: $F_{(2,123)} = 4.09$, $F_{(2,120)} = 9.37$, and $F_{(2,119)} = 12.43$ for the lower bounds, upper bounds and the range, respectively. **Figure 2** shows boxplots of the bound estimates by source uncertainty. The DMs' range estimates are most consistent with the forecasts in the imprecise condition; the median bounds under imprecision are, essentially, equal to the experts' forecasts. On the other hand, we observed the greatest reduction in range in the hybrid condition – the median range for hybrid sources is reduced by about one third – suggesting that the distinct effects of imprecision and conflict are cumulative.

Confidence and Insurance Across Sources of Uncertainty

There was a markedly different pattern of confidence and willingness-to-pay for insurance as a function of source uncertainty $F_{(2,125)} = 9.01$ and 6.51 for estimates and recommendations, respectively. DMs were most confident under imprecision and least confident under conflict. Conversely, the mean insurance bids were lowest for imprecision and highest under conflict. The two patterns were consistent, since a higher insurance bid implies lower confidence. DMs used different bidding strategies under imprecision compared to conflict and the hybrid source of uncertainty. Recall that the minimal possible bid was \$0 and the maximal was \$7.50. We considered all small bids ($\leq$\$0.25) as a rejection of insurance, very high bids ($\geq$\$7.25) as a commitment to purchasing insurance, and moderate bids $\in$ (\$0.25, 7.25) as reflecting a more nuanced conditional approach (purchase "if the price is right"). Most DMs (87%) followed the same strategy for all three sources of uncertainty, and we found no difference in the proportion of DMs who placed conditional bids (between 84% and 85%) in all conditions. DMs were more likely to reject insurance than purchase it under imprecision (12% vs. 5%). Conversely, they were more likely to commit to purchasing insurance under conflict (9% vs. 6%). In the hybrid condition the two rates were similar (8% vs. 6%). The difference between conflict and imprecision was significant, $\chi^2_{(2)} = 8.27$, $p = 0.02$ using Stuart–Maxwell test for matched categories, with 11% of DMs more likely to purchase insurance under conflict, and only 2% more likely to purchase insurance under imprecision.

Retrospective Ratings Across Sources of Uncertainty

We reverse scored "ambiguous" and "conflicting," so that high values refer to positive valence for all attributes. **Figure 3** shows the mean ratings by source uncertainty. We found no differences in the ratings of "precision," but for the other

TABLE 2 | Descriptive statistics of estimates and confidence by source uncertainty.

DV	Overall			Imprecision			Conflict			Hybrid		
	N	Mean	*SE*	*N*	Mean	*SE*	*N*	Mean	*SE*	*N*	Mean	*SE*
Best guess	390	25.40	0.22	130	25.88	0.33	130	25.10	0.42	130	25.21	0.36
Lower bound	388	13.71	0.26	129	13.22	0.43	130	13.62	0.50	129	14.29	0.42
Upper bound	382	35.59	0.30	128	36.55	0.47	128	35.38	0.56	126	34.81	0.52
Range	380	21.69	0.46	127	23.28	0.75	128	21.67	0.84	125	20.09	0.75
Confidence	390	4.02	0.07	130	4.28	0.14	130	3.78	0.12	130	3.98	0.12

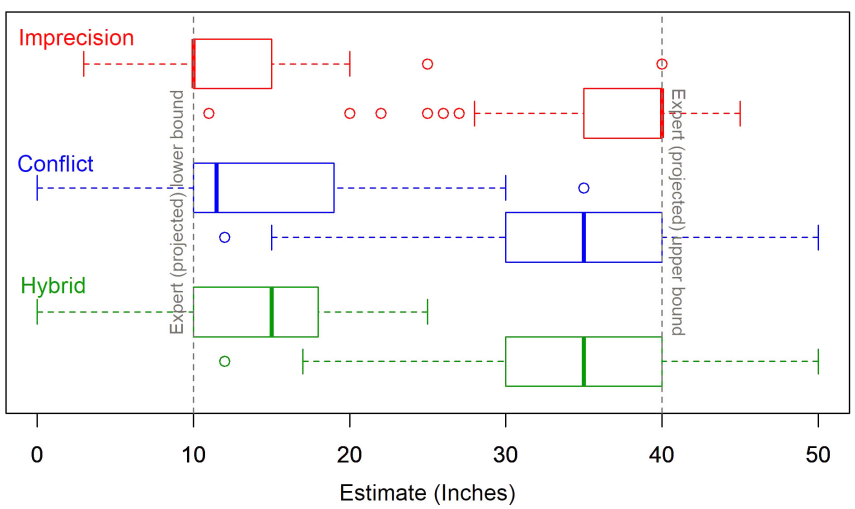


FIGURE 2 | Bound estimates by source uncertainty (Study 1).

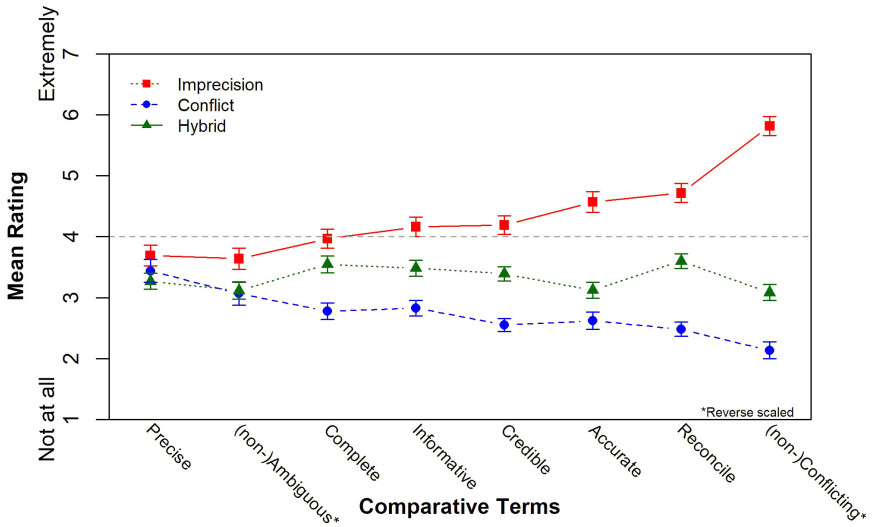


FIGURE 3 | Comparative rating means by source uncertainty (Study 1). Error bars reflect standard error of the mean for each condition.

seven attributes the differences were significant: On average, DMs rated the imprecise set highest and the conflicting set lowest. The largest difference between the three sources was observed for the rating of (non-)conflicting scale, and the smallest difference was observed for the (un-)ambiguous scale. The hybrid set was consistently rated in between the other two sets for all seven significant attributes, but its distance from the two extreme conditions varies as a function of the attribute. Mean ratings of the hybrid sets more closely resemble the mean responses under conflict for "(non-)conflicting," "(un)ambiguous," and "easy to reconcile/decide." On the other hand, the mean hybrid ratings were more like the mean ratings of the imprecise set for "completeness." The mean hybrid ratings were equidistant from imprecision and conflict for the other three attributes: "informative," "credible," and "(likely to be) accurate."

STUDY 2

In the second study, we focus on the intriguing, and previously unstudied, hybrid cases by examining the differential impact of various patterns of (interval) projections sets obtained from pairs of forecasters. The design is similar to Study 1, with some minor changes: Each DMs saw one of 30 different hybrid projection sets – involving distinct combinations of conflict and imprecision – in addition to the same set of conflicting (and precise) forecasts and the same imprecise (and agreeing) set. We also simplified the willingness-to-pay for insurance task. Rather than bidding for insurance with the BDM procedure, DMs could choose to purchase one of four levels of insurance at different pre-determined prices. Considering the results of Study 1, we did not manipulate model uncertainty. Thus, we only study source uncertainty and our hypotheses are focused on the hybrid projection sets.

We expect that DMs will react more strongly to information that reduces perceived conflict than perceived imprecision because DMs tend to be more conflict averse than imprecision averse (Smithson, 1999), as we confirmed in Study1. Following Study 1, we expect that both conflict and imprecision will contribute toward overall uncertainty, and when combined (as a hybrid or mixed source condition) their effects will be aggregated differently based on the task: (a) DMs will use a weighted mean for global preferential judgments including confidence, so we expect contributions toward overall uncertainty in the following pattern: ambiguity < hybrid < conflict, but (b) they will have combined effects for estimation tasks, so we will observe the following pattern of shifting away from the experts: ambiguity < conflict < hybrid.

We develop a systematic typology of hybrid patterns and predict that the DMs' responses will vary based on the two key factors of this classification – overlap and (a)symmetry. We expect that the type and degree of overlap between the two estimates will have a stronger influence on the global ratings than the level and nature of asymmetry between the estimates, because overlaps will drive the perceived agreement between projections. On the other hand, the degree of (a)symmetry should have a stronger influence on quantitative estimates than overlap because a large degree of asymmetry signals that averaging may not be the best method of aggregation compared to other methods (like using the median).

Methods

Participants and Design

A total of 1,084 participants completed the study online. They were recruited both via Fordham University's business school subject pool (12%) and via a Qualtrics national panel (88%). The former group received course credit, and the latter received Qualtrics' standard honorarium for completing the study. Since there were no differences between the two groups of subjects we

combine them in the analyses. In addition, 10% of all participants were randomly selected to receive an additional cash incentive of up to $14 based on their performance on one randomly selected task.

Responses were pre-screened for validity by the following pre-determined criteria to remove responses with inadequate effort: Participants must have (1) completed the survey, (2) taken at least 6 min to complete it, (3) had fewer than 15% of responses missing, and (4) straight-lined (answered identically) on at most 10 (out of 14) pages. Fewer than 14% of the responses did not meet the minimum criteria (the valid response rate did not vary by recruitment method), so we analyze a total of 937 valid responses. The sample was 50% female, the mean age was 44.4 (SD = 15.7). About 32% each self-identify as Democrats and Independents, and 28% as Republicans. Most respondents (45%) had at least a college degree, 33% had some college credit, and 22% had no college education.

Given the clear results of Study 1, we did not manipulate model uncertainty, so all participants saw forecasts from two experts and two models for each set of forecasts. The source of uncertainty was varied within-subjects, and participants saw one set of conflicting, and precise forecasts, one set of imprecise, agreeing, forecasts and one hybrid set. The unique feature of this study is that participants were randomly assigned to one of 30 conditions differentiated by 30 distinct hybrid conditions.

The various projection sets can be characterized by the type and degree of their *overlap* and (a)*symmetry*. We distinguish between four categories of overlap: (1) *intersecting* sets are partially overlapping, (2) *nested* sets feature one interval as a subset of (embedded within) the other, (3) *tangent* sets include intervals that share a common endpoint (as in Study 1), and (4) *disjoint* sets do not overlap. Let *LB1* and *LB2* be the lower bounds of the two intervals and let *UB1* and *UB2* be their corresponding upper bounds. Without loss of generality, we assume that $LB1 \leq LB2$ and $UB1 \leq UB2$ so, in other words, the first interval is lower and the second is higher. We define a measure of Degree of Overlap, *DO*, that measures the size of the interval of each set that is (dis)similar (i.e., intersecting, nested, or disjoint).

$$Degree\ of\ Overlap = DO = \begin{cases} UB1 - LB2\ if\ non\text{-}nested \\ UB2 - LB2\ if\ nested \end{cases}$$

DO is positive for intersecting and nested sets, negative for disjoint sets, and 0 for tangent sets. For example, the interesting set H01 and the nested set H06 have equal DOs, while the disjoint set H26 has a negative DO of equal magnitude (see **Figure 4**).

There are many ways to define the degree of (a)symmetry of the two sets. The basic definition we use to describe the design is based on the distance between the midpoints of the two intervals from the midpoint of all estimates. Let M be the midpoint of all four points, and m_i be the midpoint of the i'th interval (i = 1,2). Formally: $M = \frac{max(UB)-min(LB)}{2}$ and $m_i = \frac{UBi-LBi}{2}$ (i = 1, 2). We define:

$$Asymmetry = AS = |M - m_1| - |M - m_2|.$$

If the midpoints of both intervals are equidistant from M, they are considered symmetric (AS = 0). If the midpoint of the lower (upper) interval is farther from the center, M, then the set is positively (negatively) skewed. For example, set H16

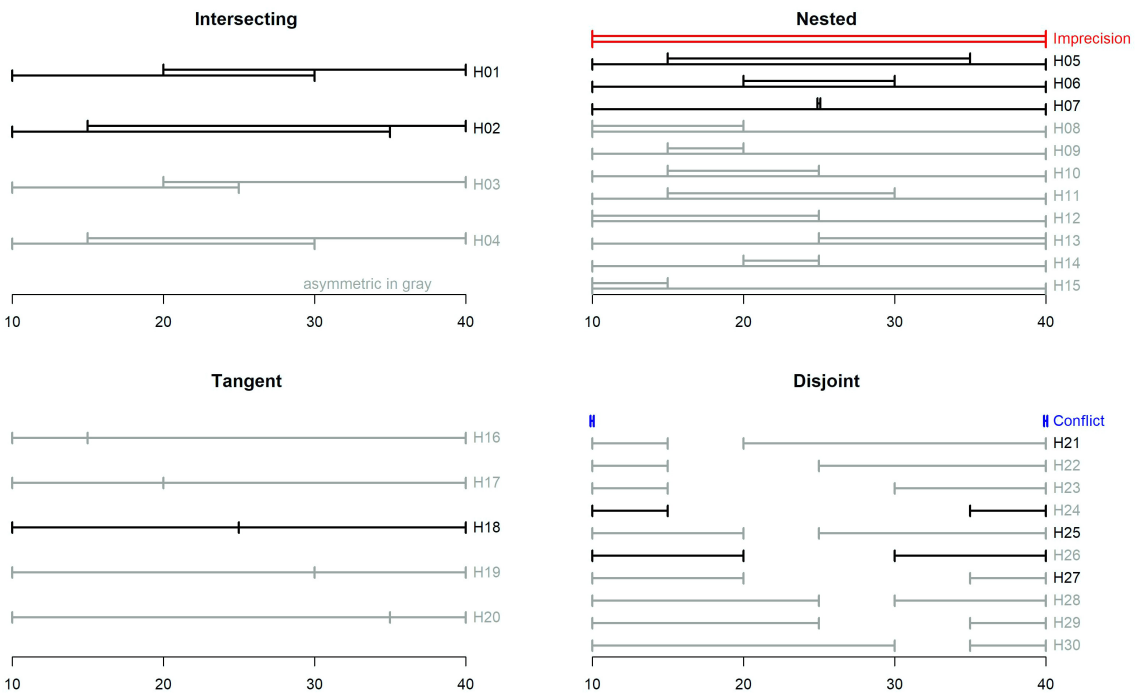


FIGURE 4 | Graphical representation of the pairs of interval forecasts used as stimuli for Study 2.

has a positive skew, and set H20 has a negative skew of equal magnitude. All forecast sets cover the same 30-inch range (from 10 to 40), so DO and AS for each set can be thought of as the breadth of that range that is overlapping (or not) and unbalanced.

Materials and Procedure

Participants were randomly assigned to one of 30 different hybrid conditions. Each included a set consisting of two interval projections that are both conflicting and imprecise to various degrees. These sets were constructed by varying the type and degree of overlap (intersecting, nesting, tangent, or disjoint) and asymmetry (symmetric or skewed), so they span the full 4×2 typology of the two factors. The complete classification is displayed in **Figure 4**.

We used the same procedure as in Study 1 with some minor modifications. Participants started by reading the informed consent and a shortened version of the background information. We used the same scenario describing how sea-level rise will affect California ports (see Supplementary Materials) showing two accompanying forecasts, but we fixed the number of experts (2) and the number of models (2). The projections were presented in a chart, and we randomized the order of presentation for each pair, so either the higher or lower projection could appear on the left or right. Participants performed the same estimation tasks and confidence ratings, but they did not provide recommendations as a government consultant. Instead, we asked them to provide a 90% probability interval, so we had a benchmark to compare the width of their range estimates.

We altered the willingness-to-pay for insurance item because the bidding procedure was time-consuming, and some participants struggled to understand it. After making their estimates, participants chose one of four levels of insurance with different levels of coverage and, of course, different costs. They could choose to (1) be uninsured, (2) pay \$1 to reduce their possible loss from \$7 to \$5, (3) pay \$2 to reduce their possible loss to \$3, or (4) pay \$3 to reduce their possible loss to \$1[3].

We also shortened the belief in CC inventory to two lists of 10 items each. In total there were two items from each subscale one of which was repeated in both lists. Beyond the six repeated items, there was one item for each subscale that was not repeated, so participants saw half of the non-duplicated items in each list. We added an item calling for self-assessment of knowledge in CC (using the same five-option scale). We used a four-item numeracy test including two items from Schwartz et al. (1997) and two items from the cognitive reflection test (Frederick, 2005) (see Supplementary Materials for scale items).

In the comparative questionnaire we dropped the rating of the trait "informative," because of its high similarity to "complete," and in the demographic questionnaire we replaced the major and year in school with highest level of education because most of the Qualtrics respondents were not students.

[3]We eliminated the quiz on the bidding procedure since we used a simple forced choice question.

Results

We ran $4 \times 2 \times 3$ mixed MANOVAs with source uncertainty (imprecision, conflict, or hybrid) as the within-subjects factor. The 30 hybrid cases were combined into four types of overlap (intersecting, nested, tangent, or disjoint) and two levels of symmetry (symmetric vs. asymmetric) and defined the between-subjects factors. We report first the results pertaining to the source of uncertainty, which replicate Study 1.

Estimates as a Function of Source of Uncertainty

We replicated the key results regarding the range estimate from Study 1. The lower and upper bounds, ranges, and the confidence ratings (see **Table 3**) varied significantly across the sources of uncertainty. DMs estimated the widest ranges under imprecision and the narrowest in the hybrid conditions. The best estimates systematically underestimated the mean of the experts' forecasts, and the deviation from the mean of the experts' forecasts was significantly greater when the two experts disagreed (conflict).

Confidence and Ratings as a Function of Source of Uncertainty

The analysis of the mean confidence and attribute ratings replicated the patterns from Study 1. DMs were most confident under imprecision, but there were no differences between conflict and the hybrid sets, and there were no significant differences across various levels of overlap or (a)symmetry. The pattern of insurance bids was similar to Study 1 with participants most likely to decline, and least likely to purchase the highest level of insurance, under imprecision. However, that difference was not significant, and fell short of a small effect, $\chi^2_{(6)} = 5.64$, $p = 0.46$, Cramer's $V = 0.03$.

Replicating the pattern from Study 1, the imprecise set was rated highest and the conflicting set was rated lowest for most attributes (there are no difference in the "precision" ratings). It is striking how stable the ratings were in their preference for imprecision and aversion to conflict for all four structures. There were main effects of source uncertainty on the ratings for the structure of overlap, but not differences between levels of asymmetry. The mean ratings of intersecting and nested hybrid sets were similar to ratings of the imprecise set and the mean ratings of the disjoint and tangent hybrid sets were

TABLE 3 | Means by source uncertainty.

Outcome	Source					
	Imprecision		Conflict		Hybrid	
	Mean	*SE*	Mean	*SE*	Mean	*SE*
Estimate (shift)	−2.18	0.28	−3.66	0.29	−2.17	0.29
Lower bound	13.56	0.24	12.67	0.24	14.07	0.25
Upper bound	31.33	0.37	29.64	0.39	29.89	0.38
Range	17.77	0.37	16.97	0.35	15.81	0.33
Confidence (est)	4.39	0.05	4.23	0.05	4.23	0.05
Insurance bid	1.19	0.04	1.24	0.04	1.23	0.04
Confidence Interval	17.47	0.44	16.99	0.44	16.11	0.40

similar to ratings of conflict. There were significant differences for the ratings of *ambiguous* and *conflicting*: nested and disjoint sets were rated the most ambiguous, and tangent and then intersecting were the least. **Figure 5** displays the mean ratings for all attributes by source of uncertainty and structure of overlap. Participants were most likely to decline insurance when viewing tangent and disjoint groups, and least likely when viewing nested and intersecting, but the effect was small and not significant, $\chi^2_{(9)} = 9.29$, $p = 0.41$, Cramer's $V = 0.03$.

The unique and novel feature of this study is the use of 30 hybrid conditions that vary along several attributes which allow us to analyze and determine if, and why, DMs are sensitive to imprecise and conflicting forecasts. Next, we discuss some of these results separately for the various dependent variables. To capture and model the subtle effects associated with various degrees of overlap and asymmetry of the 30 hybrid cases, we use regression models using the degree of each major factor and their interaction as predictors. We conducted separate regressions to predict the mean and variance of each estimate of the 30 hybrid sets allowing us to test how the key factors affect the magnitude as well as variability of the estimates. Most DMs, as expected, gave estimates close to the mean and the bounds of the set, and a small – but non-trivial – number of DMs gave greatly different responses. Focusing on the mean and variance of each group allows us to focus on the typical respondent and minimize the influence of unusual individuals.

Estimates as a Function of (A)Symmetry of the Sets

We ran regression models predicting each of the DMs' estimate using the degree of overlap, degree of asymmetry, and their interaction as predictors. The degree of asymmetry was a significant predictor of the means of all three estimates – most likely value, lower bound, and upper bound – but not for the estimated ranges, the subjective 90% probability intervals, or the confidence ratings (see **Table 4**). **Figure 6** displays the set of mean estimates by level of asymmetry. Generally, there was more shifting away from the experts' upper bound than the lower bound, and the best estimates were consistently shifted below the set mean, suggesting that DMs act as if the forecasts overestimate the "true" value. As expected, the positively skewed sets were shifted considerably toward the lower end. The symmetric set was closer to the center, but slightly shifted to the lower end. Estimates for the negatively skewed sets curbed expectations. The estimated mean was closer to the midpoint than the set mean, and the upper bounds were greatly shifted in the negative direction.

Most likely estimates were shifted further below the set mean for negatively skewed sets (−4.78, $SE = 0.62$) than both symmetric (−1.41, $SE = 0.55$; mean diff. = 3.37, 95% C.I. = [1.45, 5.29], $p < 0.001$) and positively skewed sets (−1.29, $SE = 0.41$; mean diff. = 2.76, 95% C.I. = [0.80, 4.73], $p = 0.003$)[4]. There was significantly less shifting away from the lower bound when the set had a positive skew (12.94 ($SE = 0.31$) compared to 14.91 ($SE = 0.51$), mean diff. = 1.97, 95% C.I. = [0.59, 3.35], $p = 0.002$ for symmetric and 15.51 ($SE = 0.54$), 2.58, 95% C.I. = [1.17, 3.99], $p < 0.001$ for negatively skewed sets), and significantly lower upper bound estimates for positively skewed sets (28.73, $SE = 0.55$) compared to symmetric sets (31.73 ($SE = 0.70$), mean

[4]All pair-wise comparisons based on Tukey–Kramer tests.

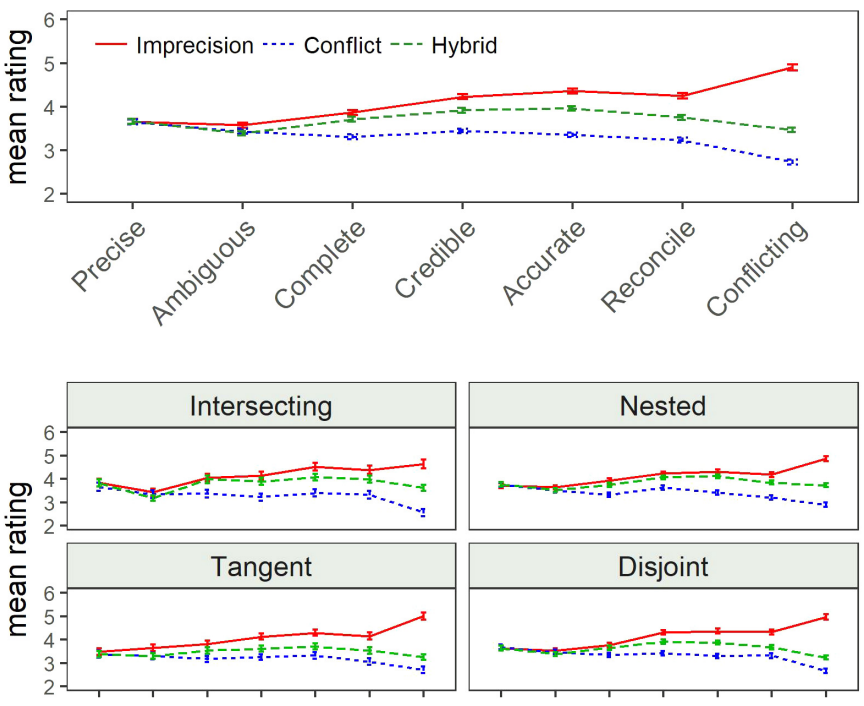


FIGURE 5 | Comparative ratings by structure of overlap.

TABLE 4 | Regression models of mean and variance of estimates by overlap and asymmetry.

Outcome	Source	Best estimate				Lower bound				Upper bound			
		B	*SE*	DW		*B*	*SE*	DW		*B*	*SE*	DW	
Mean	Intercept	−2.58	0.33			14.46	0.28			30.20	0.41		
	Overlap	0.02	0.03	0.01		−0.01	0.03	0.01		0.02	0.04	0.00	
	Asymmetry	0.27	0.06	0.22	*	−0.23	0.05	0.24	*	−0.22	0.08	0.13	*
	O'lap × Asym	0.00	0.01	0.22		0.00	0.01	0.24		0.00	0.01	0.13	
	Adj-R^2			0.39				0.43				0.17	
Variance	Intercept	81.62	6.49			58.28	6.67			131.81	11.18		
	Overlap	−0.86	0.59	0.04		−1.30	0.61	0.06	*	−0.30	1.02	0.00	
	Asymmetry	−1.53	1.24	0.02		−2.26	1.27	0.04		0.01	2.13	0.00	
	O'lap × Asym	0.13	0.16	0.08		0.23	0.16	0.15		0.03	0.27	0.00	
	Adj-R^2			0.03				0.16				−0.11	

*Significant at $p < 0.05$; DW, general dominance weight.

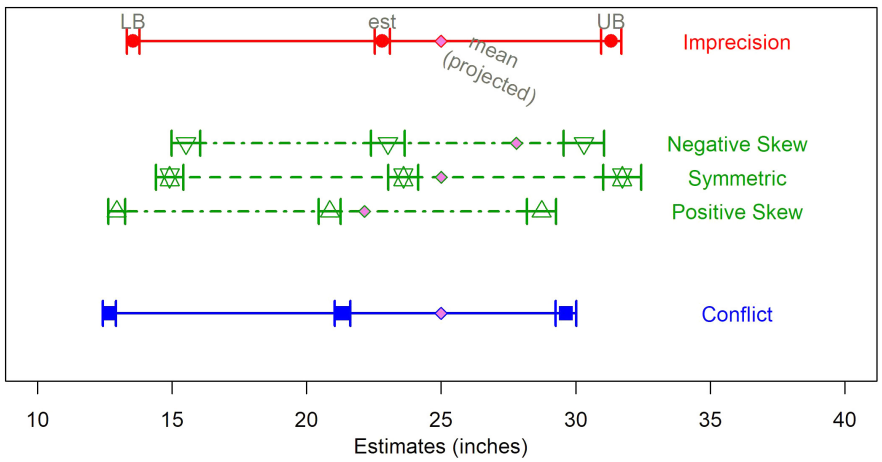


FIGURE 6 | Estimates (best, lower, and upper bounds) by level of asymmetry. Each source has a unique color, and each condition has a unique symbol. Responses to the Hybrid sets are averaged across direction of skewness. Error bars display the standard error of the mean for each condition.

diff. = 3.00, 95% C.I. = [0.86, 5.13], $p < 0.003$), but no difference from the negatively skewed sets (30.31, $SE = 0.75$).

Estimates as a Function of Overlap

The differences between the four overlap categories (intersecting, nested, tangent, or disjoint) are less pronounced. **Figure 7** displays their respective mean estimates. There was a significant difference in shift from the set mean between the categories of overlap [$F(3,926) = 3.03$, $p = 0.03$], which was driven by the difference between disjoint sets, which show the largest shift, and nested sets, with the smallest (mean diff. = 1.79, 95% C.I. = [0.01, 3.58], $p = 0.049$.

Multidimensional Scaling of the Estimates

To fully understand the response patterns across all 32 sets and the variety of measures, we ran two multidimensional scaling (MDS) analyses, one based on the estimates and the other on the post-estimation ratings.

For the first solution, we calculated the Euclidian distance between the 32 mean profiles using five responses per profile: the best estimate, lower and upper bounds and lower and upper bounds of the 90% probability intervals. A three-dimensional solution (see **Figure 8**) yields the best fit (stress = 0.04 compared to 0.23 and 0.10 for 1- and 2-dimensional solutions). The left panel colors the conditions by the degree of asymmetry (from highest positive skew in red to highest negative skew in blue). Asymmetry correlates highly with the first two dimensions (see the scatterplot matrix in the Supplementary Materials): Positively skewed sets are high, and negatively skewed sets are low, on dimensions 1 and 2.

We performed a cluster analysis on the 3-dimensional solution to help interpret it. We used hierarchical clustering with Ward linkage because it is efficient and flexible to handle both chain-like and concentric clusters. Ward's method is intuitively appealing since it minimizes the difference in sum of squares at each step in the algorithm. In the right panel of **Figure 8**, we impose the four-cluster solution that seems to be driven primarily by (a)symmetry and only to a lesser extent by overlap. One (cyan) cluster contains all negatively skewed sets (except one); a second (blue) cluster contains five (out of six) symmetric sets. The last

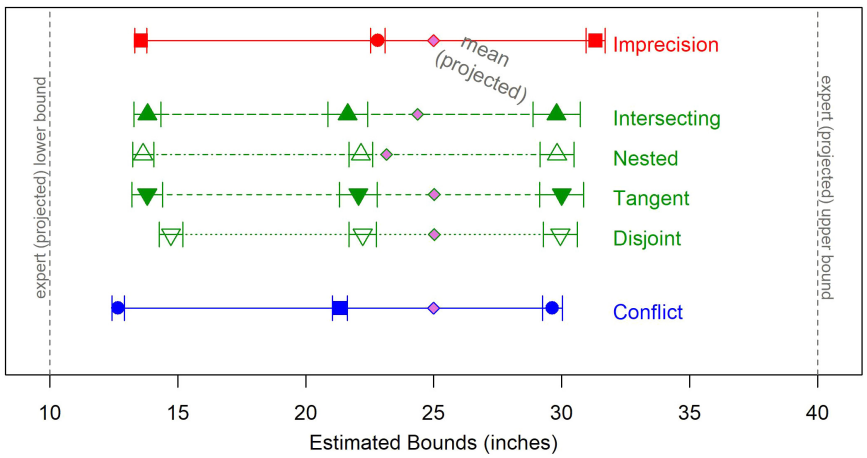


FIGURE 7 | Estimates (best, lower, and upper Bounds) by category of overlap. Each source has a unique color, and each condition has a unique symbol. Responses to the Hybrid sets are averaged across structure of overlap. Error bars display the standard error of the mean for each condition.

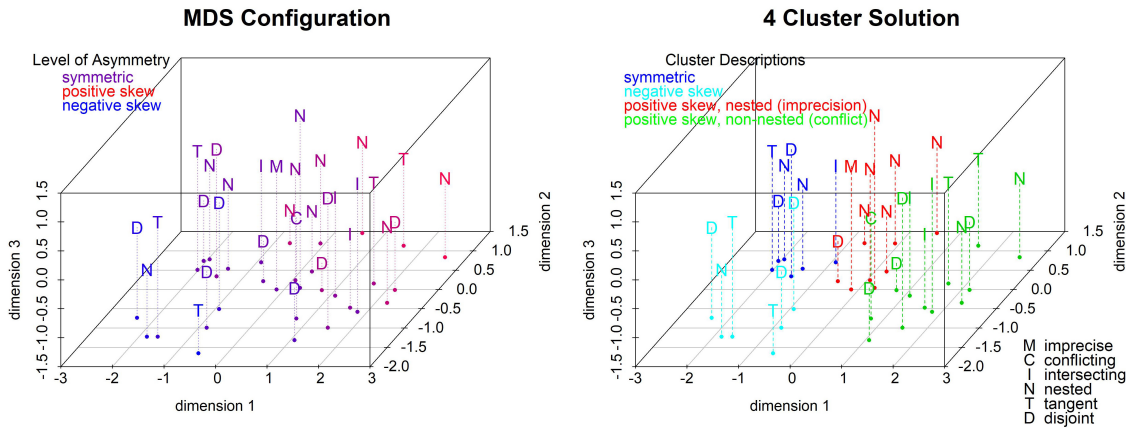


FIGURE 8 | 3-dimension MDS of all estimates by degree of asymmetry and with 4-cluster solution.

two clusters contain primarily positively skewed sets. One (red) is almost entirely (7/8) nested sets including the imprecise set and a mix of five positively skewed and three symmetric sets. The other (green) is mostly (9/12) positively skewed sets and primarily the non-nested overlap categories including three quarters of the intersecting sets, the two positively skewed tangent sets, and five disjoint sets including the conflicting set. In summary, DMs' estimates tend to vary as a function of the direction of skewness, and the nested sets lead to the most distinct estimates. This solution shows that asymmetry plays a large role in estimation, as expected, since M shifts within the projection set and the estimates shift correspondingly.

Multidimensional Scaling of the Post-estimation Ratings

We also calculated Euclidian distances between the 32 stimuli based on their seven comparative ratings: how (un-)ambiguous, (non-)conflicting, precise, credible, (likely to be) accurate, easy to reconcile, and complete each set was rated to be. The relevance of the source and type of uncertainty is apparent even for the one-dimensional solution with a stress of 0.27 (see Supplementary Materials). The conflicting case is at one end of this continuum and the imprecise set is at the other end with most hybrid sets located in-between (with only one exception at each end). The other clear result is that DMs rate disjoint and tangent sets as similar to the conflicting set (to the left end of the scale) and nested and intersecting sets as similar to the imprecise set (to the right end of the scale).

The 3-dimensional solution is the best fitting solution (stress = 0.10) and is driven by both the degree and type of overlap. The left panel of **Figure 9** colors the conditions by the degree of overlap (from highest positive overlap in red to highest negative overlap in green). Overlap correlates highly with the first two dimensions (see the scatterplot matrix in the Supplementary Materials). Sets with a positive overlap are high, and sets with a negative overlap are low, on dimensions 1 and 2. We ran a cluster analysis using Wards linkage based on the distances from the 3-dimensional MDS solution. The five-cluster solution

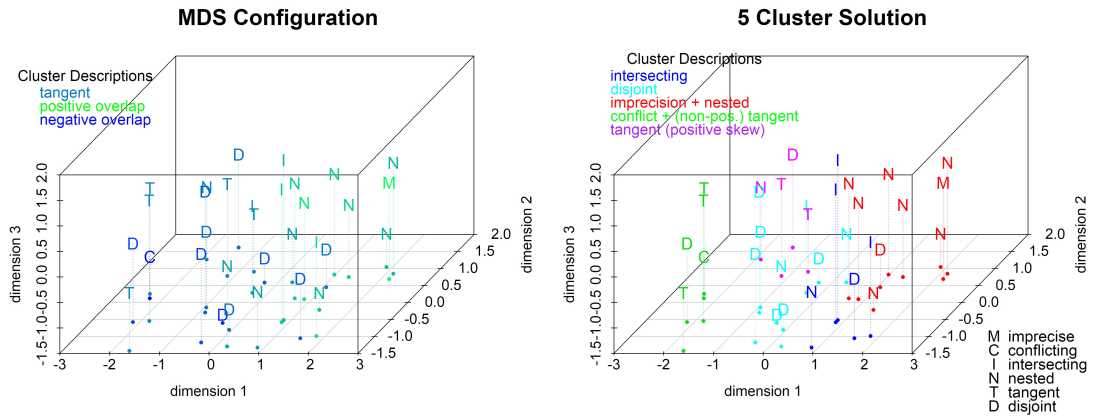


FIGURE 9 | 3-dimensional MDS of all ratings by degree of overlap and with 5-cluster solution.

(see right panel of **Figure 9**) shows that imprecision forms a cluster with the nested sets which is close to the cluster with the intersecting sets. Conflict is on the other end of the plot and the disjoint cluster is nearby. The tangent sets are divided by direction of skewness, so positively skewed form one cluster, and non-positively skewed share a cluster with conflict. In summary, tangent sets are perceived as the most conflicting followed by disjoint, and nested sets are the most agreeing followed by intersecting.

DISCUSSION

As predicted, the best estimates were, essentially, compromises between the two expert forecasts' and were unaffected by differences between the models. Most surprisingly, we found no effects of model uncertainty – neither the structural nor the judgmental component – on any of the dependent variables, but we found systematic and significant effects of source uncertainty. DMs responded to the most salient surface cues – in this case the values of the forecasts, including their relative agreement and precision, but they ignored more subtle, yet relevant, cues – the labels of the experts, models, and model parameters. These results are consistent with the system neglect hypothesis (Massey and Wu, 2005; Budescu and Yu, 2007).

There is a paradox in communicating CC information that as models get more complex, the public seems to become less sensitive to uncertainties in the models (e.g., Pidgeon and Fischhoff, 2011). Instead it appears that in domains involving deep uncertainty, such as CC, DMs are highly sensitive to the source of indeterminacy. Normatively, climate policy preferences should change according to the goal for uncertainties stemming from states of nature vs. differences between models (Drouet et al., 2015). As predicted, we observed a greater reduction in the judged range estimates (compared to the experts' original forecasts) for *conflicting point* forecasts than for *agreeing imprecise interval* forecasts. DMs show a consistent dislike of, and aversion to, conflict and react more positively to communications that reflect imprecision: (1) Imprecision resulted in the greatest consistency with the expert forecasts; (2) DMs expressed higher confidence and preferred imprecise forecasts on characteristics ranging from credibility to completeness. This supports previous findings that conflict aversion is stronger than imprecision aversion (Smithson, 1999, 2015), and implies that there is a broader dimension, such as overall perception of uncertainty, that is driven by the degree of agreement between the forecasts.

The results confirm that DMs are not universal seekers of precision, but are rather sensitive to the nature, and features, of the decision environment. They expect a certain degree of imprecision or uncertainty in climate projections, and in line with the congruence principle (Budescu and Wallsten, 1995; Olson and Budescu, 1997; Du et al., 2011), they favor forecasts that seem to capture and reflect this imprecision. In fact, using bounded estimates to express uncertainty in climate projections leads to higher belief in and concern about CC since a high degree of uncertainty is expected (MacInnis et al., unpublished).

In the presence of hybrid projections that are both imprecise and conflicting, the DMs' responses depend on the nature of the task. The joint effects of the two source uncertainties seem to lead to a compromise between the effects of imprecision and conflict for all judgmental tasks – confidence, willingness-to-pay for insurance and comparative ratings. However, we observed a combining pattern in estimation tasks – estimating the most likely value and range of possible values – where DMs displayed the least consistency with the experts in the hybrid condition. Response patterns do not reflect simple arithmetic on the endpoints. The relative agreement and configuration of the sets seem to promote differential weighting of the expert forecasts due to preferences and overall feeling about uncertainty.

These task-specific differences are consistent with the contingent weighting model (Tversky et al., 1988) and the subsequent task-goal hypothesis (Fischer et al., 1999) which state that task objectives influence response processes. We found differences in the processes used for *estimation* and *rating* tasks. In estimation tasks, DMs give range estimates closer to the endpoints under imprecision because those are the more prominent features, or focal points, while they give

range estimates closer to the middle for the hybrid set since the common point (in Study 1), which is the midpoint of the set's range, is the most prominent feature. In comparative ratings, DMs can see the hybrid set contains some features of conflict and some features of imprecision. Given that they are more averse to conflict than to imprecision (Smithson, 1999) they assign ratings that are more favorable than those of conflict, but less favorable than imprecision.

The focus of Study 2 was to develop a robust mapping of how conflict and imprecision are combined. Across the 30 different hybrid sets, reactions were a function of two key factors – structure of overlap and level of asymmetry. The results help explain why differences in cognitive processes and attention are used to respond to different goals. The size and direction of (a)symmetry of the two sets showed a stronger influence on quantitative estimates than their overlap, seemingly because asymmetry creates a tension between various measures of central tendency. Asymmetry highlights the deviation between the mean and median and makes it the prominent feature of the set. The structure and degree of overlap within a set had greater influence on preferential ratings seemingly because the (lack of) overlapping areas were the prominent set feature and altered the perception of (dis)agreement between experts. Sets with a positive overlap (nested and intersecting) were rated more similarly to imprecision indicating participants paid greater attention to the agreeing segment, and sets with a non-positive overlap (disjoint and tangent) were rated more similarly to conflict indicating participants paid greater attention to the distinct and disagreeing segments.

The two multidimensional scaling analyses confirm that estimates are driven by the *degree* of asymmetry and ratings are driven by the degree of overlap. Some pronounced response patterns within the key factors should be explored in further detail. First, heavy skewness caused the greatest bias from the set means because it creates the greatest discord between possible definitions of "center." Our hybrid sets were confined to a common range (10–40 inches), so the most skewed sets had the largest discrepancy of interval widths between projections. The large degree of bias could indicate that participants weighted the experts based on the width of their forecast. A narrower, more precise, projection seems to be associated with greater credibility. Judges perceive a tradeoff between accuracy and the informativeness of others' estimates where more narrow estimates are considered more informative, but less likely to be accurate (Yaniv and Foster, 1995). We extend these results to sets of multiple experts showing that experts providing narrower intervals are perceived as more credible and informative.

A large degree of overlap, whether an area of intersection (or positive DO) or space (negative DO), between the forecasts induced the greatest difference in ratings since it represents a larger unresolved region. A wide area of intersection suggests the experts agree, but the agreement is still imprecise. A wide area of disjointedness suggests the experts are far from agreement, and when the projections showed the most disagreement, participants rated the sets almost as conflicting and hard to resolve as pure conflict. This indicates that preferences are more complicated than following simple mathematical rules since sets with a large degree of overlap had the same range and same statistical center as corresponding sets with a small degree of overlap.

We did not observe differences based on belief in CC or political identification which, in principle, could have resulted in the discounting of projections that were inconsistent with one's political affiliation. By design, the experts and models in this study were not individualized (expert A vs. B and model GCSX vs. ESGY). Thus, the observed patterns of preference cannot be attributed to prejudices about either. Social and political attributions of the experts' motivation are a natural part of the assessment of their judgments, and it appears individuals make credibility judgments in this partisan domain even in the absence of identifying information. Moreover, recent evidence suggests attributions are a function of the educational and cognitive levels of the judges; those with lower education are more likely to attribute disagreement to incompetence, and those with higher education attribute disagreement to complexity and aleatory uncertainty (Dieckmann et al., 2015). Future work should consider the impact of these factors.

Unexpectedly we found the greatest shift away from the experts occurred when the set had a negative skew, especially for the best estimate and the lower bound. The shifting for all sets was toward lower values; best estimates were below the set mean and range estimates were narrower than the experts. This pattern is consistent with status quo bias and "system justification" – defending existing social systems – which is associated with discrediting CC (Feygina et al., 2010). Participants are less trusting of the worst-case projections, either because they do not believe the climate will continue to change at the current rate, or they tend to attribute "alarmist" motives to the forecasters who predict higher, and more threatening consequences from CC. Alternately, it implies that individuals intuit that the expected damage from CC has a specific shape that lower values are more likely than extremely high values (Lewandowsky et al., 2013). And, of course, these possibilities are not exclusive.

CONCLUSION

In two studies, we have shown that perceptions of multiple climate projections are driven by the type and degree of disagreement between them, but the judges are insensitive to the differences between the models and how they were run. Moreover, judgmental reactions to the experts are driven by how two key factors – the structure of overlap and the level of asymmetry – interact with the task at hand. It appears that previously identified uncertainties stemming from multiple sources, conflict and imprecision, are special cases of overlap and asymmetry. Perceptions of agreement require intersection and balance. While, overly precise forecasts lead to a greater perception of disagreement among experts, and a greater likelihood of the public discrediting and misinterpreting information. Future studies should build on this work by exploring how the (mis)match between the judge and various experts alters perceptions of the evidence. Further, research

should explore if overlap and asymmetry similarly impact perceptions of uncertainty in domains outside of CC.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Fordham University Institutional Review Board with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Fordham University Institutional Review Board.

AUTHOR CONTRIBUTIONS

DMB: design, data collection, analysis, and writing. DVB: design, supervision, analysis plan, and editing.

FUNDING

This work was supported by NSF Grant 1049208 (Informing Climate-Related Decisions with Earth Systems).

ACKNOWLEDGMENTS

We thank to Drs. Robert Lempert and Andrew Parker for useful conversations about the studies.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00403/full#supplementary-material

REFERENCES

Baillon, A., Cabantous, L., and Wakker, P. (2012). Aggregating imprecise or conflicting beliefs: an experimental investigation using modern ambiguity theories. *J. Risk Uncertain.* 44, 115–147. doi: 10.1007/s11166-012-9140-x

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Syst. Res. Behav. Sci.* 9, 226–232. doi: 10.1002/bs.3830090304

Bier, V. M., and Connell, B. L. (1994). Ambiguity seeking in multi-attribute decisions: effects of optimism and message framing. *J. Behav. Decis. Making* 7, 169–182. doi: 10.1002/bdm.3960070303

Budescu, D. V., Broomell, S. B., Lempert, R. J., and Keller, K. (2014a). Aided and unaided decisions with imprecise probabilities in the domain of losses. *EURO J. Decis. Process.* 2, 31–62. doi: 10.1007/s40070-013-0023-4

Budescu, D. V., Kuhn, K. M., Kramer, K. M., and Johnson, T. R. (2002). Modeling certainty equivalents for imprecise gambles. *Organ. Behav. Hum. Decis. Process.* 88, 748–768. doi: 10.1016/S0749-5978(02)00014-6

Budescu, D. V., Por, H. H., Broomell, S. B., and Smithson, M. (2014b). The interpretation of IPCC probabilistic statements around the world. *Nat. Clim. Change* 4, 508–512. doi: 10.1038/nclimate2194

Budescu, D. V., and Wallsten, T. S. (1995). Processing linguistic probabilities: general principles and empirical evidence. *Psychol. Learn. Motiv.* 32, 275–318. doi: 10.1016/S0079-7421(08)60313-8

Budescu, D. V., and Yu, H. T. (2007). Aggregation of opinions based on correlated cues and advisors. *J. Behav. Decis. Making* 20, 153–177. doi: 10.1002/bdm.547

Cabantous, L., Hilton, D., Kunreuther, H., and Michel-Kerjan, E. (2011). Is imprecise knowledge better than conflicting expertise? Evidence from insurers' decisions in the United States. *J. Risk Uncertain.* 42, 211–232. doi: 10.1007/s11166-011-9117-1

Casey, J. T., and Scholz, J. T. (1991). Boundary effects of vague risk information on taxpayer decisions. *Organ. Behav. Hum. Decis. Process.* 50, 360–394. doi: 10.1016/0749-5978(91)90027-Q

Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *Int. J. Forecast.* 5, 559–583. doi: 10.1016/0169-2070(89)90012-5

Cokely, E. T., Feltz, A., Ghazal, S., Allan, J. N., Petrova, D., and Garcia-Retamero, R. (2018). "Decision making skill: from intelligence to numeracy and expertise," in *Cambridge Handbook of Expertise and Expert Performance*, 2nd Edn, eds K. A. Ericsson, R. R. Hoffman, A. Kozbelt and A. M. Williams (New York, NY: Cambridge University Press).

Dieckmann, N. F., Johnson, B. B., Gregory, R., Mayorga, M., Han, P. K., and Slovic, P. (2017). Public perceptions of expert disagreement: bias and incompetence or a complex and random world? *Public Underst. Sci.* 26, 325–338. doi: 10.1177/0963662515603271

Dieckmann, N. F., Mauro, R., and Slovic, P. (2010). The effects of presenting imprecise probabilities in intelligence forecasts. *Risk Anal.* 30, 987–1001. doi: 10.1111/j.1539-6924.2010.01384.x

Dieckmann, N. F., Peters, E., and Gregory, R. (2015). At home on the range? Lay interpretations of numerical uncertainty ranges. *Risk Anal.* 35, 1281–1295. doi: 10.1111/risa.12358

Dieckmann, N. F., Peters, E., Gregory, R., and Tusler, M. (2012). Making sense of uncertainty: advantages and disadvantages of providing an evaluative structure. *J. Risk Res.* 15, 717–735. doi: 10.1080/13669877.2012.666760

Ding, D., Maibach, E. W., Zhao, X., Roser-Renouf, C., and Leiserowitz, A. (2011). Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nat. Clim. Change* 1, 462–466. doi: 10.1038/nclimate1295

Doran, P. T., and Zimmerman, M. K. (2009). *Examining the Scientific Consensus on Climate Change, Eos, Transactions*, Vol. 90. Washington, DC: American Geophysical Union, 22–23. doi: 10.1029/2009EO030002

Drouet, L., Bosetti, V., and Tavoni, M. (2015). Selection of climate policies under the uncertainties in the Fifth Assessment Report of the IPCC. *Nat. Clim. Change* 5, 937–940. doi: 10.1038/nclimate2721

Du, N., and Budescu, D. V. (2005). The effects of imprecise probabilities and outcomes in evaluating investment options. *Manag. Sci.* 51, 1791–1803. doi: 10.1287/mnsc.1050.0428

Du, N., Budescu, D. V., Shelly, M. K., and Omer, T. C. (2011). The appeal of vague financial forecasts. *Organ. Behav. Hum. Decis. Process.* 114, 179–189. doi: 10.1016/j.obhdp.2010.10.005

Einhorn, H. J., and Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychol. Rev.* 92, 433–461. doi: 10.1037/0033-295X.92.4.433

Einhorn, H. J., and Hogarth, R. M. (1988). "Decision making under ambiguity: a note," in *Risk, Decision and Rationality Theory and Decision Library (Series B: Mathematical and Statistical Methods)*, Vol. 9, ed. B. R. Munier (Netherlands: Springer), 327–336.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Q. J. Econ.* 75, 643–669. doi: 10.2307/1884324

Erev, I., and Cohen, B. L. (1990). Verbal versus numerical probabilities: efficiency, biases, and the preference paradox. *Organ. Behav. Hum. Decis. Process.* 45, 1–18. doi: 10.1016/0749-5978(90)90002-Q

Feygina, I., Jost, J. T., and Goldsmith, R. E. (2010). System justification, the denial of global warming, and the possibility of "system-sanctioned change". *Pers. Soc. Psychol. Bull.* 36, 326–338. doi: 10.1177/0146167209351435

Fischer, G. W., Carmon, Z., Ariely, D., and Zauberman, G. (1999). Goal-based construction of preferences: task goals and the prominence effect. *Manag. Sci.* 45, 1057–1075. doi: 10.1287/mnsc.45.8.1057

Fischhoff, B., and Davis, A. L. (2014). Communicating scientific uncertainty. *Proc. Natl. Acad. Sci. U.S.A.* 111(Suppl. 4), 13664–13671. doi: 10.1073/pnas.1317504111

Frederick, S. (2005). Cognitive reflection and decision making. *J. Econ. Perspect.* 19, 25–42. doi: 10.1257/089533005775196732

Galesic, M., Kause, A., and Gaissmaier, W. (2016). A sampling framework for uncertainty in individual environmental decisions. *Topics Cogn. Sci.* 8, 242–258. doi: 10.1111/tops.12172

González-Vallejo, C., Bonazzi, A., and Shapiro, A. J. (1996). Effects of vague probabilities and of vague payoffs on preference: a model comparison analysis. *J. Math. Psychol.* 40, 130–140. doi: 10.1006/jmps.1996.0012

Hammond, K. R. (1996). *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. New York, NY: Oxford University Press.

Heath, Y., and Gifford, R. (2006). Free-market ideology and environmental degradation: the case of belief in global climate change. *Environ. Behav.* 38, 48–71. doi: 10.1177/0013916505277998

Heath, C., and Tversky, A. (1991). Preference and belief: ambiguity and competence in choice under uncertainty. *J. Risk Uncertain.* 4, 5–28. doi: 10.1007/BF00057884

Hogarth, R. M., and Einhorn, H. J. (1990). Venture theory: a model of decision weights. *Manag. Sci.* 36, 780–803. doi: 10.1287/mnsc.36.7.780

Holyoak, K. J., and Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *J. Exp. Psychol. Gen.* 128, 3–31. doi: 10.1037/0096-3445.128.1.3

Joslyn, S. L., and LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *J. Exp. Psychol. Appl.* 18, 126–140. doi: 10.1037/a0025185

Kahan, D. M., Jenkins-Smith, H., and Braman, D. (2011). Cultural cognition of scientific consensus. *J. Risk Res.* 14, 147–174. doi: 10.1080/13669877.2010.511246

Kramer, K., and Budescu, D. V. (2004). "Exploring Ellsberg's paradox in vague-vague cases," in *Experimental Business Research, 3*, eds R. Zwick and A. Rapoport (Netherlands: Kluwer Academic Publishers), 131–154.

Kuhn, K. M. (1997). Communicating uncertainty: framing effects on responses to vague probabilities. *Organ. Behav. Hum. Decis. Process.* 71, 55–83. doi: 10.1006/obhd.1997.2715

Kuhn, K. M., and Budescu, D. V. (1996). The relative importance of probabilities, outcomes, and vagueness in hazard risk decisions. *Organ. Behav. Hum. Decis. Process.* 68, 301–317. doi: 10.1006/obhd.1996.0107

Kunda, Z. (1990). The case for motivated reasoning. *Psychol. Bull.* 108, 480–498. doi: 10.1037/0033-2909.108.3.480

Larrick, R. P., and Soll, J. B. (2006). Intuitions about combining opinions: misappreciation of the averaging principle. *Manag. Sci.* 52, 111–127. doi: 10.1287/mnsc.1050.0459

Lewandowsky, S., Gignac, G. E., and Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nat. Clim. Chang.* 3, 399–404. doi: 10.1038/nclimate1720

Lewandowsky, S., Risbey, J. S., Smithson, M., Newell, B. R., and Hunter, J. (2014). Scientific uncertainty and climate change: part I. Uncertainty and unabated emissions. *Clim. Change* 124, 21–37. doi: 10.1007/s10584-014-1082-7

Massey, C., and Wu, G. (2005). Detecting regime shifts: the causes of under-and overreaction. *Manag. Sci.* 51, 932–947. doi: 10.1287/mnsc.1050.0386

Morgan, M. G., and Keith, D. W. (1995). Subjective judgments by climate experts. *Environ. Sci. Technol.* 29, 468A–476A. doi: 10.1021/es00010a753

N.C. Coastal Resources Commission's Science Panel on Coastal Hazards (2010). *North Carolina Sea-level Rise Assessment Report March 2010*. Available at: http://www.sealevel.info/NC_Sea-Level_Rise_Assessment_Report_2010--CRC_Science_Panel.pdf

Newell, B. R., and Pitman, A. J. (2010). The psychology of global warming: improving the fit between the science and the message. *Bull. Am. Meteorol. Soc.* 91, 1003–1014. doi: 10.1175/2010BAMS2957.1

Olson, M. J., and Budescu, D. V. (1997). Patterns of preference for numerical and verbal probabilities. *J. Behav. Decis. Making* 10, 117–131. doi: 10.1002/(SICI)1099-0771(199706)10:2<117::AID-BDM251>3.0.CO;2-7

Pidgeon, N., and Fischhoff, B. (2011). The role of social and decision sciences in communicating uncertain climate risks. *Nat. Clim. Chang.* 1, 35–41. doi: 10.1038/nclimate1080

Russo, J. E., and Yong, K. (2011). The distortion of information to support an emerging evaluation of risk. *J. Econom.* 162, 132–139. doi: 10.1016/j.jeconom.2010.07.004

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. Inter. Med.* 127, 966–972. doi: 10.7326/0003-4819-127-11-199712010-00003

Shanteau, J. (2000). "Why do experts disagree," in *Risk Behaviour and Risk Management in Business Life*, eds B. Green, R. Cressy, F. Delmar, T. Eisenberg, B. Howcraft, M. Lewis, et al. (Dordrecht: Kluwer Academic Press), 186–196.

Siceloff, B. (2014). *While the Seas Rise in the Outer Banks and Elsewhere in NC, Science Treads Water. The News & Observer*. Available at: http://www.newsobserver.com/news/politics-government/state-politics/article10298660.html

Smithson, M. (1999). Conflict aversion: preference for ambiguity vs conflict in sources and evidence. *Organ. Behav. Hum. Decis. Process.* 79, 179–198. doi: 10.1006/obhd.1999.2844

Smithson, M. (2015). Probability judgments under ambiguity and conflict. *Front. Psychol.* 6:674. doi: 10.3389/fpsyg.2015.00674

Tversky, A., Sattath, S., and Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychol. Rev.* 95, 371–384. doi: 10.1037/0033-295X.95.3.371

Wallsten, T. S., and Budescu, D. V. (1995). A review of human linguistic probability processing: general principles and empirical evidence. *Knowl. Eng. Rev.* 10, 43–62. doi: 10.1017/S0269888900007256

Wallsten, T. S., Budescu, D. V., Erev, I., and Diederich, A. (1997). Evaluating and combining subjective probability estimates. *J. Behav. Decis. Making* 10, 243–268. doi: 10.1002/(SICI)1099-0771(199709)10:3<243::AID-BDM268>3.0.CO;2-M

Wallsten, T. S., Budescu, D. V., Zwick, R., and Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bull. Psychon. Soc.* 31, 135–138. doi: 10.3758/BF03334162

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., and Peters, E. (2013). Development and testing of an abbreviated numeracy scale: a Rasch analysis approach. *J. Behav. Decis. Making* 26, 198–212. doi: 10.1002/bdm.1751

Yaniv, I., and Foster, D. P. (1995). Graininess of judgment under uncertainty: an accuracy-informativeness trade-off. *J. Exp. Psychol. Gen.* 124, 424–432. doi: 10.1037/0096-3445.124.4.424

Yaniv, I., and Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organ. Behav. Hum. Decis. Process.* 103, 104–120. doi: 10.1016/j.obhdp.2006.05.006

Zehr, S. C. (2016). Public representations of scientific uncertainty about global climate change. *Public Underst. Sci.* 9, 85–103. doi: 10.1088/0963-6625/9/2/301

Zickfeld, K., Morgan, M. G., Frame, D. J., and Keith, D. W. (2010). Expert judgments about transient climate response to alternative future trajectories of radiative forcing. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12451–12456. doi: 10.1073/pnas.0908906107

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

BOOK REVIEW
published: 15 May 2018
doi: 10.3389/fpsyg.2018.00730

Book Review: Handbook of the Economics of Risk and Uncertainty

Shabnam Mousavi [1,2]*

[1] Johns Hopkins Carey Business School, Washington, DC, United States, [2] Max Planck Institute for Human Development, Berlin, Germany

Keywords: risk, uncertainty, probability, decision theory, economics

A Book Review on

Handbook of the Economics of Risk and Uncertainty

Mark J. Machina and W. Kip Viscusi (North Holland: Elsevier), 2014. ISBN: 978-0-444-53685-3.

Consider the future as a product of interplay between the states of the nature on one hand and our choices on the other. Perceivably, we can make a particular future come true if we can specify possible outcomes of choices and their relative likelihood. Needless to say, we shall always choose the best option. Economists employ mathematics and logic to make this conviction concrete. Addressing these issues, the *Handbook of the Economics of Risk and Uncertainty* consists of two masterfully crafted prefaces and 14 chapters written by leading economists in theory, empirical, and experimental economics. Below I highlight some central concepts that are examined from different perspectives in many (though not all) chapters. Corresponding chapters and sections in the handbook that discuss each topic are indicated inside parentheses.

Bet on what you believe in. This adage was made concrete by the seventeenth-century representation of beliefs in possible lottery outcomes, artfully complemented three centuries later with the operationalization of the inference of beliefs from observed choices. The latter enabled specifying prior beliefs about future prospects, which was missing from the original Bayesian approach to updating beliefs based on new information (1). Not only could beliefs be represented as specifiable probability distributions, but also the best value or maximum utility could be calculated for rational players whose well-behaved preference rankings were capable of being captured in utility functions. Under risk, where all prospects and their probabilities can be objectively specified, rationality is mainly reflected in the *independence axiom*, which holds that the introduction of a third option, z, should not alter an initial preference order between two existing options, x and y: $x \succcurlyeq y \rightarrow \alpha x + (1-\alpha)\, z \succcurlyeq \alpha y + (1-\alpha)\, z$. Allais famously produced lottery choices that violate this essential axiom, launching an ongoing line of literature (2).

Moving from risk to situations of uncertainty, probabilities of prospects need to be subjectively assessed. Here the consistency requirement of rationality is preserved by Savage's *sure-thing principle*, which assigns a premium to a given prospect equal to the expected value of the lottery, tantamount to rational risk aversion. However, Ellsberg's famous experiment revealed that not all uncertainties can be captured by subjective probability assignments—giving rise to the concept of *ambiguity* and much follow-up work (2.6, 13, 14.4). Probabilities can be classified according to the distinction not only between objective and subjective but also between aleatory and epistemological. When risk is not objectively known, it can be assessed subjectively, even if it is essentially knowable. On the other hand, economic risk corresponds to the aleatory category of probabilities arising from relative frequencies in repeated trials, whereas uncertainty corresponds to the epistemological category of probabilities, as in degrees of belief. Both meanings seem to lose operational relevance when unknown prospects are involved. This third category of unknowns is referred to as *ignorance* and is material for future research (Preface 2).

OPEN ACCESS

Edited and reviewed by:
David R. Mandel, Defence Research and Development Canada, Canada

***Correspondence:**
Shabnam Mousavi
shabnam@jhu.edu

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

Received: *26 January 2018*
Accepted: *26 April 2018*
Published: *15 May 2018*

Citation:
Mousavi S (2018) Book Review: Handbook of the Economics of Risk and Uncertainty. Front. Psychol. 9:730. doi: 10.3389/fpsyg.2018.00730

Actions do not affect probabilities. This is the main flavor of expected utility calculations. Von Neumann and Morgenstern's (vNM) expected utility theory (EUT) concerns the formation of strategies, mixed and otherwise, for noncooperative, zero-sum situations with no pure equilibrium when uncertainty is objectified as risk (1.2, 3.3). Maximizing a utility function that satisfies the three axioms of vNM—namely, completeness, transitivity, and continuity—is equivalent to choosing the best possible prospect, which by definition is the most preferred option. Savage's contributions to decision theory came in two phases. First, his subjective probability theory provided a framework for constructing relative likelihoods of prospects without preference ordering. Second, his subsequent axiomatic approach to choice under uncertainty defined necessary and sufficient criteria for the joint existence and uniqueness of utility and probability for choices with deterministic consequences in static situations, thereby extending vNM utilities to the subjective level (1.3, 14.1). Further extensions of this idea to dynamic situations by others (2.5, 14.2) dictated that only *naïve* agents who change taste at every stage or *myopic* agents who overlook future stages violate intertemporal consistency, whereas *resolute* agents keep executing the initial plan despite changes in preferences and *sophisticated* agents plan by backward induction based on perfect foresight of their future taste developments, hence acting in a consistent manner along a dynamic path. Thus, resolute and sophisticated agents are rational agents for whom time does not affect planned actions.

The conception of expected utilities can be traced back to the 18th century when, with the introduction of diminishing marginal utility, Daniel Bernoulli remedied the inadequacy of expected value maximization, posed for one by the St. Petersburg paradox. Nonetheless, until the mid-twentieth century, that is, prior to EUT, economists remained focused on analysis of valuation in terms of simple mean-variance (M-V) utility functions, such as $V(\sigma, \mu) = \mu - \lambda.\sigma^2$, that rank the agents' preference over random returns (3). This ranking, which is independent of all higher moments, remains to date the main tenet of asset pricing, where the tradeoff between risk and return can be optimized for an investor with given preferences. In model building, these preferences were assumed as given. In the laboratory, risk preferences are elicited in one of three ways (4, 7.2): the proportion of investment in risky versus safe assets in a portfolio, the point at which subjects switch from a risky to a safe gamble on a given menu, and the named selling or buying price for a gamble, which reveals certainty equivalents. The EU ranking coincides with the M-V ranking for normal distribution and generally in the case of a CARA (constant absolute risk aversion) utility function (3.6). Otherwise, when higher moments are significant, such as in skewed distributions, econometrics methods provide nonlinear representations for assessment of risk preferences (4.3).

In sum, the contributors to this handbook view rational decision making as static or dynamic and model it in combination with deterministic, risky, or uncertain consequences. The impetus of the majority of arguments lies in experiments conducted mainly by economists. This collection is deeply rooted in theoretical and axiomatic conceptualizations of decision making under risk and uncertainty with a sprinkling of the psychological studies of heuristics (4.7). This handbook is most useful for cognitive scientists and psychologists who want to learn about the background details of what economists explored and entertained that are now known as central notions of behavioral economics, presented in psychology terminology such as risk aversion, domain of gain versus loss, and reference point. These very concepts, only in different terms, can be traced back to the joint work of Friedman and Savage from 1948 and the subsequent investigations by Harry Markowitz, who observed: "Generally people avoid symmetric bets. This suggests that the curve falls faster to the left of the origin than it rises to the right of the origin."

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

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

SYSTEMATIC REVIEW
published: 11 October 2017
doi: 10.3389/fpsyg.2017.01785

Meta-Analytic Evidence for a Reversal Learning Effect on the Iowa Gambling Task in Older Adults

Rita Pasion[1,2], Ana R. Gonçalves[1], Carina Fernandes[1,3], Fernando Ferreira-Santos[1], Fernando Barbosa[1] and João Marques-Teixeira[1]*

[1] Laboratory of Neuropsychophysiology, Faculty of Psychology and Educational Sciences, University of Porto, Porto, Portugal, [2] Católica Porto Business School, Universidade Católica Portuguesa, Porto, Portugal, [3] Faculty of Medicine, University of Porto, Porto, Portugal

Iowa Gambling Task (IGT) is one of the most widely used tools to assess economic decision-making. However, the research tradition on aging and the Iowa Gambling Task (IGT) has been mainly focused on the overall performance of older adults in relation to younger or clinical groups, remaining unclear whether older adults are capable of learning along the task. We conducted a meta-analysis to examine older adults' decision-making on the IGT, to test the effects of aging on reversal learning (45 studies) and to provide normative data on total and block net scores (55 studies). From the accumulated empirical evidence, we found an average total net score of 7.55 (±25.9). We also observed a significant reversal learning effect along the blocks of the IGT, indicating that older adults inhibit the prepotent response toward immediately attractive options associated with high losses, in favor of initially less attractive options associated with long-run profit. During block 1, decisions of older adults led to a negative gambling net score, reflecting the expected initial pattern of risk-taking. However, the shift toward more safe options occurred between block 2 (small-to-medium effect size) and blocks 3, 4, 5 (medium-to-large effect size). These main findings highlight that older adults are able to move from the initial uncertainty, when the possible outcomes are unknown, to decisions based on risk, when the outcomes are learned and may be used to guide future adaptive decision-making.

Keywords: Iowa Gambling Task, decision-making, risk, uncertainty, aging, older adults, neuropsychology

OPEN ACCESS

Edited by:
Nathan Dieckmann, Oregon Health & Science University, United States

Reviewed by:
Yang Jiang, University of Kentucky, United States
David Copeland, University of Nevada, Las Vegas, United States

****Correspondence:***
Rita Pasion
ritapasion@gmail.com

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

Received: *27 July 2017*
Accepted: *26 September 2017*
Published: *11 October 2017*

Citation:
Pasion R, Gonçalves AR, Fernandes C, Ferreira-Santos F, Barbosa F and Marques-Teixeira J (2017) Meta-Analytic Evidence for a Reversal Learning Effect on the Iowa Gambling Task in Older Adults. Front. Psychol. 8:1785. doi: 10.3389/fpsyg.2017.01785

INTRODUCTION

Decision-making is a fundamental process in everyday life and subject to major changes over the lifespan. According to a recent meta-analysis, early adolescents show a pattern of risk-seeking behavior compared to mid-late adolescents, despite similar performances in decision-making between children and adolescents (Defoe et al., 2015). Moreover, adolescents were found to be more risk-seeking in tasks with immediate feedback compared to adults (Defoe et al., 2015).

A meta-analysis in healthy older adults was further conducted by Mata et al. (2011), while differentiating participants' decision-making performance in contexts of uncertainty and risk. Uncertainty refers to circumstances in which the probabilities of the possible outcomes are unknown, while in decisions under risk the outcomes and probabilities are given in advance (Ellsberg, 1961; Kahneman and Tversky, 1979; Mata et al., 2011). Older adults seem to

engage more often in disadvantageous decisions than younger adults, but only under uncertainty (Mata et al., 2011). Under risk, younger and older adults showed similar patterns of decision-making (Mata et al., 2011). Interestingly, older adults engage less in risky activities compared to younger adults, and are more responsive to warnings about potential risks (Rolison et al., 2016). The aging-related reduction in risk-taking seems to occur steeply for financial and recreational decisions, but smoothly for ethical and health-related decisions (Rolison et al., 2013). It seems that, under risk, older adults respond to threat levels with increased cautiousness (Rolison et al., 2013, 2016), but the threat level may be difficult to identify without previous information, as occurs in decision-making under uncertainty (Mata et al., 2011).

The processes and mechanisms that may explain the abovementioned age differences in decisions under risk and uncertainty are still poorly understood. The Iowa Gambling Task (IGT) (Bechara et al., 1994) may advance our knowledge on the differential patterns of performance under risk and uncertainty, since the decisions along this task are expected to move from uncertainty to risk (Brand et al., 2007).

In the IGT, participants are asked to choose a card from four different decks to win as much money as possible. While performing the task it is expected that participants learn to discriminate advantageous (Decks C and D) from disadvantageous decks (Decks A and B). However, learning during the IGT requires more than just adjusting behavior as a function of reliable feedback signaling long-term correct and incorrect responses. Considering that high rewards in the IGT are included in the disadvantageous decks, the prepotent response is initially oriented to the decks that are also associated with increased losses (Kovalchik and Allman, 2006). Adaptive behavior requires the inhibition of the prepotent response, as participants learn to forego the high monetary rewards (immediately attractive options that are also associated with high losses) in favor of the low to moderate monetary rewards (initially less attractive options that are associated with reduced losses and long-run profit). The shift in the prepotent response during the learning process is conceptualized as the reversal learning effect (Kovalchik and Allman, 2006). The difference between the number of disadvantageous (Decks A and B) and advantageous choices (Decks C and D) is considered a Gambling Index—the total net score (Bechara et al., 1994)—that captures the reversal learning effect and the adaptive course of action. Since the implicit feedback in the first half of the task is considered a close correlate of the uncertainty experienced in real-life (Bechara et al., 1994; Verdejo-Garcia et al., 2006; Bechara, 2007; Brevers et al., 2013), the reversal learning effect may unveil the moment in which participants learn the advantageous strategy. Brand et al. (2007) reported that only the last trials of IGT were correlated with the performance under risk, supporting that decks probabilities are learned along the task.

The starting point to conceptualize the shift from uncertainty to risk is grounded on Damásio's Somatic Marker Hypothesis (Damasio, 1994). Damasio (1994) proposed that the affective signals generated from the match between choices and associated outcomes guide subsequent decisions, by biasing the decision to the options associated with positive affective states. The affective states are detected by the limbic system, particularly, the amygdala. During the first trials, the limbic areas trigger the affective values of gains and losses, and generate the automatic somatic states (primary inducers) (Bechara et al., 2003; Bechara and Damasio, 2005). Interestingly, levels of uncertainty are positively associated with amygdala activation, suggesting that this structure is recruited to detect relevant information when the probabilities are unknown (Hsu et al., 2005).

An affective executive system—the hot executive functioning (EF)—also accounts to detect monetary rewards and losses under uncertainty. The hot EF is defined as the set of abilities that regulate emotional awareness, impulsive reactions and goal achievement, by integrating emotional, affective, and visceral processes (Miyake et al., 2000; Zelazo and Müller, 2002; Séguin et al., 2007; Brevers et al., 2013). The incentive saliency measured under uncertainty is monitored by the hot executive functioning to signal the best outcomes. Then, the information from primary inducers is accommodated in the memory systems (Bechara et al., 2003; Bechara and Damasio, 2005), and the hot EF assists the cold EF in an integrated decision process (Zelazo and Müller, 2002; Séguin et al., 2007; Brevers et al., 2013). The cold EF is conceptualized as the cognitive determinant of risk and gains, updating and maintaining the information in working memory (Zelazo and Müller, 2002; Séguin et al., 2007; Brevers et al., 2013). The hot and cold EF interplay is a critical process to plan the necessary changes to future choices (Zelazo and Müller, 2002; Brevers et al., 2013). The ventromedial prefrontal cortex (vmPFC) will be critical to guide future adaptive decision-making, mediating the secondary inducers—somatic states generated by the recall of emotional events (Bechara et al., 2000, 2003; Bechara and Damasio, 2005). The vmPFC activation is associated with global IGT performance (Northoff et al., 2006; Lawrence et al., 2009; Li et al., 2010) and, interestingly, with performance in the final trials (Northoff et al., 2006), suggesting that the vmPFC becomes less dependent on amygdala-driven autonomic responses at the end of the task (Bechara et al., 2003).

Functional age-related changes in brain areas implicated in decision-making [e.g., insula and anterior cingulate cortex (Good et al., 2001), superior temporal sulcus (Sowell et al., 2003), dorsal and ventral striatum (Raz et al., 2005; Walhovd et al., 2005), prefrontal (West, 2000), and orbitofrontal cortex (Resnick et al., 2007)] suggest that older adults exhibit less resources to decide adaptively. EF, mainly dependent on prefrontal areas, are particularly vulnerable to age-related cognitive decline (Best et al., 2009). From the 7th decade of life, a detrimental effect is found in several executive domains, such as response inhibition, planning, and set shifting (Best et al., 2009), that are important functions to reversal learning. Older adults tend to make more perseverative errors, which indicate an inability to plan future behavior in function of previous feedback and a failure to inhibit an activated response pattern that as proven to be disadvantageous.

A reversal learning effect in older players performing the IGT is not detected when compared to younger players (Kovalchik and Allman, 2006). Kovalchik and Allman (2006) proposed that the lack of an initial preference in older adults compromises the subsequent process of reversal learning. Decision-making and

reversal learning seem, therefore, to become inoperative with age, suggesting that a random selection strategy may be guiding decision-making in elderly (Kovalchik and Allman, 2006).

Steingroever et al. (2013) also proposed that IGT performance on healthy groups are characterized by slow learning processes, and 100 trials are not sufficient to learn to discriminate the safe over the risky options. The infrequent occurrence of losses in decks B and D provides little information to learn that deck B should be avoided. Moreover, with the exception of the deck A, the remaining decks seem to have too similar outcomes (Steingroever et al., 2013). Participants fail, therefore, to distinguish bad from good decks, failing to progress from an initial stage of exploration to a later stage of exploitation. The limitations of the learning processes expected to occur during the IGT may be particularly observed in older groups, since this group have increased difficulty in discriminating negative from positive outcomes in reinforcement learning tasks. The reduced Feedback-Related Negativity (FRN) amplitude was found to be similar after losses and gains, suggesting a decreased focus of the monitoring system in classifying the outcomes according to task-specific goals (Hämmerer et al., 2011). In fact, older adults need more trials to identify the option more likely to be rewarded, particularly when differences in reward likelihood between choices are small (Hämmerer et al., 2011).

The revised literature documents that older adults show detrimental changes in decision-making brain-related areas that may compromise critical functions as EF, reversal and reinforcement learning that are critical processes to decision-making under uncertainty. However, and considering that decisions under risk are similar (Mata et al., 2011) or even improved compared to younger adults (Rolison et al., 2013, 2016), it remains unclear whether older adults are capable of learning adaptive strategies and move from uncertainty toward risk, that is, to integrate affective automatic responses in memory and rational analytical systems that facilitate future adaptive decision-making.

The current meta-analysis aims to address this gap in the literature. The IGT includes both initial stages of exploration (decisions under uncertainty) and later stages of exploitation (decisions under risk) (Brand et al., 2007; Steingroever et al., 2013), providing a comprehensive analysis of decision-making. This allows extending Mata et al.'s (2011) results, which were obtained with tasks assessing decision-making under risk and under uncertainty independently. Also, Mata et al.'s (2011) conclusions are retrieved from studies with a between-group design (older vs. younger groups), from which we cannot infer directly that older adults are not capable of learning.

For this purpose, we have meta-analyzed the performance of older adults along the IGT blocks. The within-subject design of our meta-analysis allow monitoring the participants' performance along the task and to isolate the reversal learning effect. The analysis of the shift from uncertainty to risk is of great importance, since older adults' difficulties in decision-making appear to be restricted to uncertainty (Mata et al., 2011). We hypothesize that the contrasting pattern of performance under risk and uncertainty is explained by the lack of a reversal learning effect in older adults (Kovalchik and Allman, 2006). The reversal learning effect is required to perform adaptively in tasks under uncertainty, and subsequently to move to a context of decision-making under risk, in which the task contingencies have been learned and may guide future adaptive decisions.

This hypothesis constitutes an innovative approach to the IGT, since the results are typically analyzed in terms of the total net score, disregarding the dynamics of learning that occurs within the task. Finally, we also provide normative data on older adults' performance from the literature reviewed, namely a group reference criterion to compare individual values of IGT total and block net scores.

METHODS

The current meta-analysis followed the PRISMA Statement guidelines for reporting systematic reviews and meta-analyses (Moher et al., 2009).

Eligibility Criteria

The focus of the systematic search was studies that assessed economic decision-making processes in older adults with the IGT.

As inclusion criteria, the studies had to: (1) describe empirical results; (2) report the Bechara et al.'s (1994) original version of the IGT, in its manual or computerized versions; (3) include a sample of healthy older adults (mean age $\geq$ to 55 years old and standard deviation $<$ to 10). Mean age criteria was based on Denburg's et al. (2005) cut off, and standard deviation criteria was defined to avoid samples with a large interval of age.

Studies were excluded if: (4) none of the parameters of the current review (total and block IGT net scores) were reported; and (5) contained overlapping results.

To avoid publication bias, we considered unpublished results, but none were retained after the application of inclusion/exclusion criteria.

Study Selection

PubMed, EBSCOhost (Academic Search Complete, PsycARTICLES, Psychology and Behavioral Sciences Collection), and Web of Knowledge databases were used to identify papers published since the first administration of IGT (Bechara et al., 1994) (1994–September 2016).

The search expression, limited to titles and abstracts in English, was (neurodegenerative OR Alzheimer OR Parkinson OR Huntington OR dementia OR "mild cognitive impairment" OR "frontotemporal dementia" OR ageing OR aging OR "older adults" OR elderly) AND ("Iowa Gambling Task"). Neurodegenerative disorders were included in the search expression to identify papers using healthy adults as controls.

The selection of the studies included the following steps: (1) combination of search results from different databases and removal of duplicates; (2) assessment of inclusion criteria by two independent raters (RP, CF), considering the abstract and full text. Disagreements were resolved by consensus; (3) reference lists were screened to identify additional relevant papers; (4) authors were contacted to provide missing information; (5) papers with missing or repeated data were excluded.

Data Collection and Variables Extracted

During the assessment of inclusion criteria, the inter-rater agreement Cohen's kappa was used to compare the agreement between the researchers, revealing an almost perfect agreement ($K = 0.95$).

A standardized coding form was then developed to systematically collect the main parameters of analysis. This process was conducted by two independent researchers (RP and ARG).

The extracted variables from the included studies were: final sample size (n), gender (n females), age (mean and standard deviation), years of education (mean and standard deviation), task administration (computerized or manual), compensation (none, fixed, proportional to the performance), total and blocks net score (means and standard deviations).

Data Analysis

The quantitative results obtained from total and block net scores were used to achieve our main goals.

To compute the normative data, the standard errors of the mean extracted from the figures were first converted to standard deviation by multiplying the standard error of the mean by the square root of the sample size (Higgins and Green, 2011). Pooled means (M_{pooled}) and pooled standard deviations (SD_{pooled}) were then calculated for each study. These pooled parameters compose a single unit of analysis in which larger sample sizes are proportionally represented by a greater effect on the overall estimate, which improves the estimate precision and allows to compare independent sample estimates.

To explore the effects of aging on reversal learning, we computed the magnitude of the effect size from the difference between block net scores, always in reference to the baseline (block 1; B1). All analyses were conducted on Comprehensive Meta-Analysis software (3.0; Biostat, U.S.A).

The meta-analytic methods were performed in accordance to a within-subject design, as the data for the same participants was entered for more than one condition, introducing statistical dependence between the conditions. The work on statistical methods for meta-analysis has been more focused on independent sample sizes, whereas repeated measures received more limited attention. The use of the same methods to calculate independent and dependent effects is not recommended, as it introduces significant estimate bias on within-subject designs (Dunlap et al., 1996; Morris and DeShon, 2002). The calculation of the within-subject effect size is dependent on the value of the correlation between conditions, in addition to the means and standard deviations of each condition. However, this value is rarely provided in research reports. Indeed, none of the reviewed studies reported this correlation, because the main interest was to test group differences between the older group and the younger or clinical groups. This issue has a critical relevance to the current meta-analysis, as the magnitude of the effect size depends on whether the correlations between the conditions are smaller, larger, or equal to 0.5 (Morris and DeShon, 2002; Ferreira-Santos, in press).

Given the issues outlined above, we opted to impute the estimated correlation from the databases provided by the authors, where the correlation values between all blocks were available (seven samples). Correlations were pooled by using the weighted average of Z-transformed coefficient coefficients, which was then transformed back into a correlation coefficient via Fisher's Z inverse transformation (Silver and Dunlap, 1987). Because the distribution of Z is approximately normal, this method tends to be less biased than a simple arithmetic average, which distribution becomes negatively skewed as the correlation is larger than zero, particularly when including small samples (Silver and Dunlap, 1987).

All the seven samples used to impute the correlation coefficient revealed a low correlation between blocks: $r = -0.107$ to 0.219 (B2-B1); $r = -0.039$ to 0.379 (B3-B1); $r = -0.155$ to 0.198 (B4-B1); $r = -0.024$ to 0.277 (B5-B1). This resulted in an imputed r value of 0.007 in B2-B1, of 0.086 in B3-B1, of 0.018 in B4-B1, and of 0.068 in B5-B1. For the seven samples where the r value was available, the original value was maintained in accordance to the performances between blocks. To assess whether the variation in the correlation value would modify the reported effect size, a sensitivity analysis using a range of plausible correlations was conducted using moderated (0.50) and high (0.80) correlations.

From the imputed correlation coefficients, we calculated the Hedge's g. This method prevents the overestimation of the absolute value of the effect size parameter in studies with small samples (Hedges, 1981), as frequently observed for Cohen's d (Cohen, 1988).

High scores on g indicate a positive net outcome (i.e., better decisions on later decks when compared to the first deck), while negative values are associated with negative outcomes and disadvantageous behavior.

Heterogeneity Analysis

The heterogeneity analysis allows testing the consistency of results across included studies. Statistical heterogeneity between studies is considered inevitable, since methodological diversity always occurs (Higgins and Green, 2011).

The variability between studies, that is, differences in effect sizes that are caused by other factors than chance (sampling error), was tested using the Q test (Cochran, 1954) and I^2 (Higgins and Green, 2011). The significance of Q indicates the presence of heterogeneity, while the I^2 describes the percentage of the variability in effect estimates that is due to heterogeneity. Heterogeneity was present in the current meta-analysis, suggesting that there is in fact more than one true effect sizes at the population-level. Considering this, we may not assume that individual effect sizes are measures of a single population effect size (fixed-effects models). The alternative is to incorporate the heterogeneity in the random-effect models, where individual estimates are measures of a distribution of possible population-level effect sizes (Field, 2001; Schmidt et al., 2009; Higgins and Green, 2011). Providing a hyperparameter of the population distribution, random-effect models allow us to generalize the findings to the population, whereas inferences based on fixed-effects models are restricted to the set of the studies reviewed (Schmidt et al., 2009).

Moderation Analysis

To further explore the factors that may be accounting to the heterogeneity in results, we performed a moderation analysis.

Sample (proportion of females, age and years of education) and task variables (administration and compensation) are variables systematically identified in the literature and thought to modulate the performance in the IGT (for a review see Fernie and Tunney, 2006; Areias et al., 2013). However, the lack of variability in task administration and compensation variables, in addition to restricted information from several variables of interest, conditioned the assessment of moderation effects of these variables. Consequently, these variables were only used to better characterize the studies in which the use of the normative data may be particularly relevant.

Regarding the sample characteristics, age, years of education, and percentage of females were considered as continuous moderator variables. Independent meta-regressions across block performance were conducted to test the learning effect, when moderated by age, years of education, and the percentage of females in the sample.

RESULTS

Study Inclusion

Detailed information on the study selection process is described in the PRISMA Flow Diagram (**Figure 1**).

A total of 403 non-duplicated articles were found and 10 studies were added by cross-reference check.

In nine studies, it was not possible to assess inclusion criteria. Authors were then contacted and asked to provide more detailed information about the mean age and respective standard deviation of the samples. Responses were not obtained for one study (response rate = 88.9%) that was, therefore, excluded based on inclusion criterion 3.

Twenty-one studies did not meet the inclusion criterion 1, 45 the inclusion criterion 2, and 279 studies the inclusion criterion 3.

For the 68 eligible papers, only three studies reported all the required information to test our main hypothesis. For the remaining studies, the authors were contacted. Data was no longer available for five studies, but additional information was provided for 25 studies. Of note, Caselli et al. (2011) kindly sent to us a larger dataset of the published study. Lamar also provided us with a more recent database from Visagan et al.'s (2012) study. This latter paper only reports IGT performance for a younger group, but the authors kindly authorized us to report the performance of the healthy older adults collected at the time. For the studies with no response or with no information available, the total and block net scores were extracted from the graphical illustrations using Engauge Digitizer software (V9.8, https://markummitchell.github.io/engauge-digitizer/). However, 14 articles did not contain the required information and were removed from the analysis (exclusion criterion 4).

The contact with the authors and the overlap of the outcome measures also allowed us to identify repeated data across studies (exclusion criterion 5). One study was removed.

Fifty-three articles (55 cells) were retained to calculate the normative data for the IGT performance in aging. A subset of 44 studies (45 cells) was used to test the effect of aging on the reversal learning effect. All the studies were published between 2002 and 2016[1].

Sample

The data from 1977 older adults (55% female) were used to calculate normative data on the IGT performance (**Table 1**). The mean age of the sample was 68.2 years and the mean years of formal education was 13.2.

Task

Seventy-six percent of the studies used the computerized version of the IGT. The remaining studies did not report the procedure associated with task administration (manual vs. computerized), which led to the exclusion of this parameter from the moderator analysis.

None of the included studies rewarded participants based on the IGT performances. Of the five studies reporting payment to participants, only four compensated the participants. Considering the few data points available, compensation was removed from moderator analysis.

Normative Data

The samples from the included studies had an age range of 55–79 years old. Normative data for the total and blocks net scores is presented at **Table 1**.

Reversal Learning Effect

Figure 2 provides a graphical illustration of the reversal learning effect. There is a significant small-to-medium effect size considering the difference of block 2 performance relative to block 1, $g = 0.48$, 95% CI [0.37, 0.58], $Z = 8.50$, $p < 0.001$. Forest plot of B2-B1 is displayed at **Figure 3**. The reduced consistency between studies suggests that block 2 still corresponds to an exploratory stage of learning, that is, the trials performed up to this point were not sufficient to learn to discriminate task contingencies.

From the block 3 onward, a gradual increase in a medium-to-large effect size is found in relation to block 1: B3-B1, $g = 0.70$, 95% CI [0.56, 0.84], $Z = 9.70$, $p < 0.001$; B4-B1, $g = 0.73$, 95% CI [0.58, 0.89], $Z = 9.17$, $p < 0.001$; B5-B1, $g = 0.74$, 95% CI [0.58, 0.89], $Z = 9.26$, $p < 0.001$. Forest plot of B5-B1 comparison is displayed at **Figure 4**. The increase in effect size magnitude suggests that, in later blocks, studies systematically report that older adults learn to discriminate advantageous from disadvantageous decks.

The remaining forest plots (B3-B1 and B4-B1) may be found in Figures 5, 6 (Supplementary Information).

[1]Despite systematic search procedures, meta-analyses often evidence publication bias since studies reporting significant differences and large effects, are more likely to be published than studies that report non-significant differences (e.g., Dickersin, 2015). To assess the overestimation bias in the reported effect sizes we calculated the Egger's (1997) regression intercept for each block comparison. Egger's intercept was not significant in B2-B1 ($B = 1.04$, $p = 0.172$). However, a significant intercept was found in B3-B1 ($B = 2.97$, $p < 0.001$), B4-B1 ($B = 3.59$, $p < 0.001$), and B5-B1 ($B = 3.58$, $p < 0.001$), suggesting the existence of publication bias.

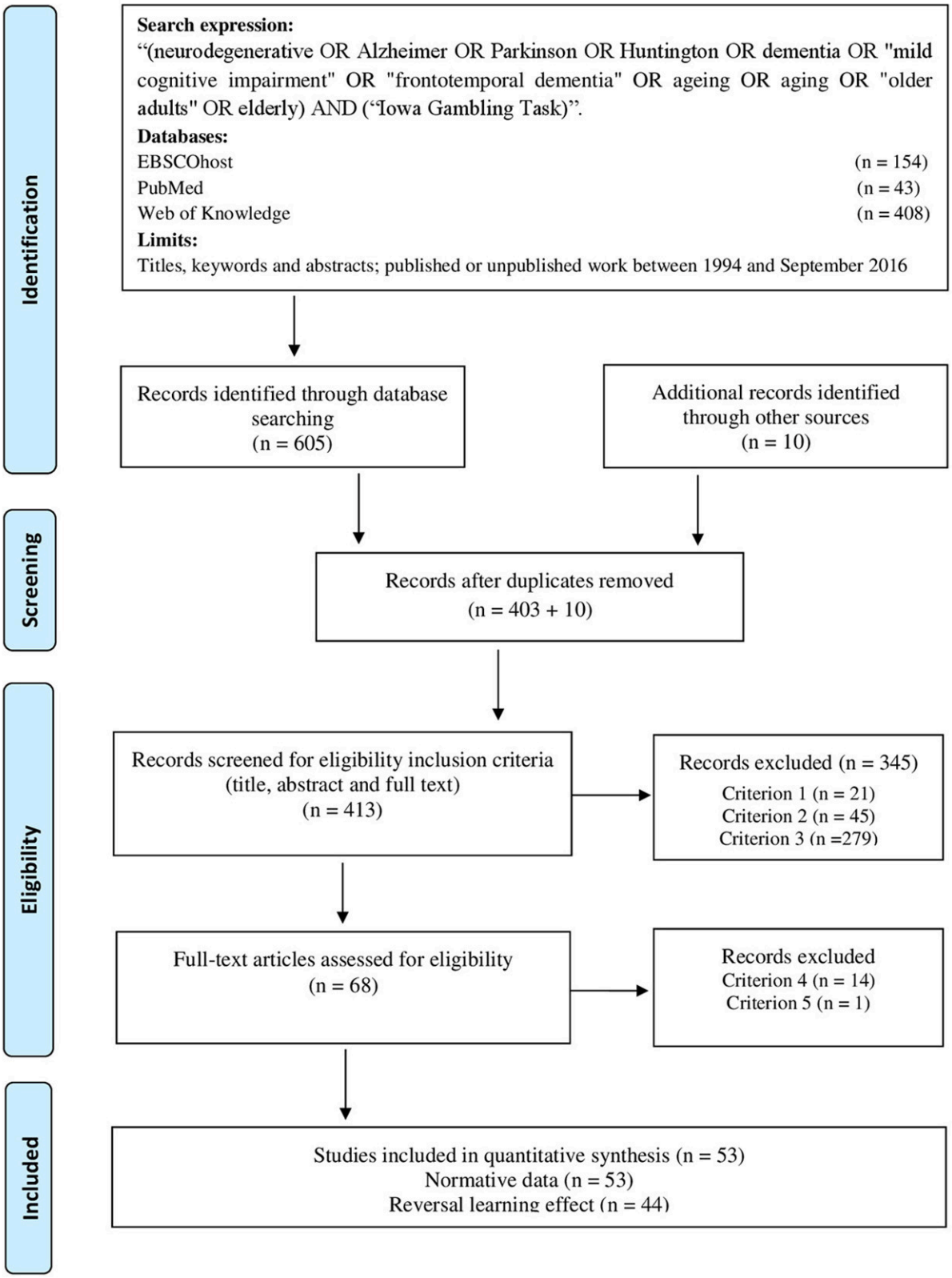


FIGURE 1 | PRISMA flow diagram.

TABLE 1 | Study characteristics and normative total and block net scores for the 55–79 age range.

	Sample				Task		Iowa Gambling Task net score												
	Size		Age	Education	Computorized	Compensation	Block 1		Block 2		Block 3		Block 4		Block 5		Total net score		
	n	*Female*	*M*	*M*			*M* (*k* = 51)	*SD* (*k* = 45)	*M* (*k* = 51)	*M* (*k* = 45)	*SD* (*k* = 31)	*SD* (*k* = 45)	*M* (*k* = 51)	*SD* (*k* = 45)	*M* (*k* = 51)	*SD* (*k* = 45)	*M* (*k* = 43)	*SD* (*k* = 31)	Net outcome
MacPherson et al., 2002	30	15	69.9	12.4		–	−4.11	8.23	−0.76	10.1	−0.18	5.19	−1.01	10.18	0.39	8.74	–	–	–
Rosi et al., 2016	72	43	68.0	14.6	Yes	–	−1.92	8.99	3.78	7.98	3.78	8.22	4.75	8.99	4.00	10.4	14.39	28.9	↑
Auzou et al., 2014	18	7	68.0	1.40		–	−1.56	6.38	−1.56	5.14	−3.00	7.61	0.00	7.27	0.22	10.3	−5.24	24.9	↓
Ibarretxe-Bilbao et al., 2009	24	8	57.8	13.0	Yes	–	−1.17	3.80	1.04	4.29	9.15	5.62	12.1	6.45	14.9	6.54	36.5	22.7	↑
Mapelli et al., 2014	15	4	60.7	11.4	Yes	–	−5.27	10.3	0.03	10.2	2.98	10.1	7.20	10.27	9.89	10.9	–	–	–
Poletti et al., 2010	25	11	65.4	9.30	Yes	–	−2.33	2.61	−0.27	3.01	1.33	3.26	3.20	1.97	4.20	4.54	6.10	6.8	↑
Torralva et al., 2009	14	7	65.5	13.9	Yes	–	1.56	7.48	0.91	6.38	5.67	6.36	5.32	6.38	8.89	8.59	–	–	–
Icellioglu, 2015	30	15	66.3	9.54	Yes	No	−4.48	6.04	−0.72	5.21	3.87	3.23	1.57	5.44	1.84	6.79	3.45	12.8	↑
Schiebener and Brand, 2016	42	26	67.5	–		–	−2.50	5.41	1.45	6.65	1.14	8.17	2.10	8.79	3.74	9.50	−0.19	25.1	↓
Carvalho et al., 2012	40	30	67.4	14.7	Yes	–	−0.15	5.44	0.95	5.54	2.45	6.26	−0.15	6.67	0.55	8.37	3.93	21.6	↑
Euteneuer et al., 2009	23	11	64.4	11.7	Yes	–	−3.65	3.70	1.48	5.53	3.00	7.15	2.17	6.95	2.52	7.94	5.48	22.7	↑
Kobayakawa et al., 2008	22	9	67.6	14.4	Yes	–	1.09	3.58	0.00	4.94	0.64	5.10	3.64	4.392	1.73	6.60	4.90	2.60	↑
Bakos et al., 2008	10	9	62.0	14.1	Yes	–	0.01	–	1.23	–	2.69	–	0.84	–	4.92	–	15.4	23.1	↑
Czernecki et al., 2002	28	10	58.1	12.7	Yes	–	−4.79	3.71	0.14	6.53	0.79	8.06	5.00	9.97	6.43	10.9	7.60	4.20	↑
Balconi et al., 2014a	42	24	57.0	13.6	Yes	–	0.76	14.2	1.82	14.6	7.75	14.8	8.98	14.1	12.5	13.8	–	–	–
Caselli et al., 2011	110	77	63.4	16.4		–	−1.39	9.77	4.45	8.89	5.40	9.74	5.98	8.99	4.15	10.6	18.6	26.8	↑
Evens et al., 2015	32	9	65.3	10.7	Yes	–	−0.09	7.33	−1.44	6.61	−2.31	5.74	−1.22	6.08	−2.91	6.74	−7.31	20.8	↓
Kloeters et al., 2013	28	12	64.2	13.7	Yes	–	−1.10	8.40	2.80	7.10	5.90	9.50	7.30	9.30	4.00	8.70	18.9	28.9	↑
Smart and Krawitz, 2015	25	10	69.9	16.9	Yes	–	−0.34	0.76	2.81	7.68	7.50	16.9	6.98	15.1	3.98	9.08	–	–	–
Bertoux et al., 2013	30	15	67.2	10.7	Yes	–	−2.40	3.90	−0.13	5.70	0.80	5.80	−0.20	8.00	−1.33	7.70	−3.2	22.9	↓
Isella et al., 2008	40	22	65.4	8.70		–	−0.96	24.0	1.03	39.4	2.72	52.5	3.59	59.2	0.81	38.0	7.1	19.6	↑
Zamarian et al., 2008	52	34	69.3	–	Yes	–	−4.00	5.57	2.15	6.16	5.27	7.79	5.54	8.76	7.35	8.28	16.4	20.7	↑
Delazer et al., 2012	45	23	63.9	11.8	Yes	–	−3.44	11.5	0.56	19.2	4.79	8.26	7.90	9.74	8.77	8.84	18.4	27.54	↑
Balconi et al., 2014b	39	17	56.4	13.9	Yes	–	−0.11	0.44	1.01	0.62	3.60	0.87	5.90	0.75	8.02	0.69	–	–	–
Gleichgerrcht et al., 2012	14	7	65.5	13.9		–	1.50	7.60	0.80	6.40	5.60	6.60	5.20	6.70	8.80	8.50	21.9	19.9	↑
Torralva et al., 2007	10	6	63.5	13.5	Yes	–	1.31	8.38	0.87	6.86	5.69	7.30	5.64	7.51	8.98	9.56	22.9	–	↑
Baena et al., 2010	39	–	69.9	–	Yes	–	−0.26	13.4	2.31	14.4	3.13	16.8	1.44	17.8	2.10	16.9	1.74	15.9	↑
Sasai et al., 2012	34	13	64.0	13.0	Yes	–	−0.48	3.28	0.51	1.91	1.99	1.77	1.36	1.37	3.55	2.87	6.90	16.6	↑
Ottaviani and Vandone, 2011	114	91	59.1	–	Yes	–	−3.84	6.09	−1.95	6.98	−1.47	8.95	−1.95	8.92	−0.23	9.49	−9.44	30.3	↓
Cardoso et al., 2014	18	14	59.3	12.1	Yes	–	−0.44	5.20	3.44	6.31	7.44	7.41	6.00	9.10	5.89	9.03	23.0	19.2	↑
Schneider and Parente, 2006	40	27	68.0	–	Yes	–	−6.16	–	−2.67	–	−2.44	–	11.6	–	−0.05		–	–	–
Wagner et al., 2009	27	24	69.6	10.4	Yes	–	−0.64	–	−1.38	–	−0.57	–	−1.02	–	0.62		−17.9	17.3	↓
Manes et al., 2011	14	7	65.5	13.9		–	1.56	7.18	0.94	5.85	5.54	6.68	5.36	5.68	8.79	8.52	21.9	19.9	↑

(Continued)

TABLE 1 | Continued

	Sample				Task		Iowa Gambling Task net score												
	Size		Age	Education	Computorized	Compensation	block 1		block 2		block 3		block 4		block 5		Total net score		
	n	*Female*	*M*	*M*			*M* (*k* = 51)	*SD* (*k* = 45)	*M* (*k* = 51)	*M* (*k* = 45)	*SD* (*k* = 31)	*SD* (*k* = 45)	*M* (*k* = 51)	*SD* (*k* = 45)	*M* (*k* = 51)	*SD* (*k* = 45)	*M* (*k* = 43)	*SD* (*k* = 31)	Net outcome
Visagan et al., 2012	35	–	67.3	–	Yes	–	−3.71	7.56	1.31	6.51	1.31	8.98	2.51	9.53	3.60	10.3	5.29	31.35	↑
Buelow et al., 2014	13	7	69.6	15.9	Yes	–	−0.92	6.25	1.85	6.71	2.31	9.38	6.31	8.79	6.77	11.7	15.9	27.92	↑
Damholdt et al., 2011	33	15	68.1	–	Yes	–	−0.61	6.13	1.76	7.79	0.30	7.13	−0.79	8.15	−0.48	9.76	0.18	25.1	↑
Delazer et al., 2009	20	17	71.3	10.3	Yes	–	−4.79	4.57	1.63	5.69	5.72	7.22	7.75	7.14	8.07	8.26	–	–	–
Fein et al., 2007	52	34	73.7	16.2	Yes	Yes	−1.77	11.9	1.09	9.41	3.84	11.8	6.15	11.5	5.98	14.5	15.0	33.4	↑
Bayard et al., 2015	20	11	73.5	11.1	Yes	–	−2.20	3.72	1.10	4.38	2.20	7.70	3.30	8.19	4.50	9.49	8.90	26.1	↑
Wyart et al., 2016	43	17	79.3	20	Yes	–	−2.19	6.13	−0.23	8.27	2.51	8.10	4.56	11.5	4.65	13.1	9.30	30.4	↑
Evans-Roberts and Turnbull, 2010	10	6	71.5	13.6		–	–	–	–	–	–	–	–	–	–	–	29.8	20.9	↑
Bakos et al., 2008	10	9	79.6	14.1	Yes	–	−0.99	–	−3.46	–	−1.86	–	−2.26	–	−6.87	–	−14.6	6.40	↓
Caselli et al., 2011	76	54	75.5	16.1	–	–	−2.63	8.69	3.84	7.88	3.13	9.71	2.34	10.9	2.28	10.6	8.95	29.1	↑
McGovern et al., 2014	36	–	71.6	16.4	Yes	–	−3.58	12.7	2.26	8.81	4.68	9.21	3.46	9.62	0.66	9.45	–	–	–
Perretta et al., 2005	19	8	72.6	14.3	–	–	5.10	–	5.62	–	6.63	–	6.62	–	6.05	–	–	–	–
Alexopoulos et al., 2015	30	0	72.8	16.3	–	Yes	−2.60	9.87	2.60	9.01	5.67	9.61	5.60	9.79	0.87	10.9	12.1	29.4	↑
Pagonabarraga et al., 2007	31	15	70.2	9.90	Yes	–	2.00	–	0.00	–	1.00	–	2.00	–	4.00	–	6.10	17.0	↑
Zamarian et al., 2011	22	11	76.0	9.50	Yes	–	−3.55	6.01	3.91	7.29	4.36	10.6	5.00	9.23	4.55	8.86	14.3	28.8	↑
Sinz et al., 2008	22	17	75.2	10.0	Yes	–	−5.00	4.89	1.64	4.60	6.09	6.03	7.36	7.18	7.00	8.57	17.1	17.5	↑
Delpero et al., 2015	27	14	73.5	9.90	Yes	–	−1.29	2.68	1.00	2.58	4.57	6.48	3.29	6.91	3.07	6.02	–	–	–
Hot et al., 2014	32	19	75.7	–	Yes	–	−2.13	2.00	−0.56	1.26	2.56	2.81	5.19	3.24	5.94	2.88	2.2	5.89	↑
Denburg et al., 2005	40	20	70.7	–	–	Yes	−3.19	6.86	−0.74	7.21	1.03	11.9	−0.86	12.3	1.83	11.2	–	–	–
Huang et al., 2015	65	47	75.3	–	–	Yes	–	–	–	–	–	–	–	–	–	–	−3.48	28.3	↓
Shivapour et al., 2012	116	73	73.6	15.9	Yes	–	–	–	–	–	–	–	–	–	–	–	12.1	38.0	↑
Denburg et al., 2009	79	50	74.0	15.8	Yes	–	–	–	–	–	–	–	–	–	–	–	5.35	–	↑
	1,977	1,081	68.2	13.2	42	4													
Normative data							−1.97	8.59	1.14	10.2	2.90	12.1	3.69	13.2	3.73	11.4	7.55	25.9	

Sensitivity Analysis

The sensitivity analysis did not reveal major alterations in the reported effect sizes using either a moderate 0.05 correlation coefficient (B2-B1; $g = 0.47$, 95% CI [0.36, 0.58], $Z = 8.43$, $p < 0.001$; B3-B1, $g = 0.72$, 95% CI [0.60, 0.86], $Z = 9.91$, $p < 0.001$; B4-B1, $g = 0.77$, 95% CI [0.61, 0.93], $Z = 9.58$, $p < 0.001$; B5-B1, $g = 0.77$, 95% CI [0.61, 0.93], $Z = 9.62$, $p < 0.001$), or a 0.08 correlation coefficient (B2-B1, $g = 0.46$, 95% CI [0.35, 0.60], $Z = 8.36$, $p < 0.001$; B3-B1, $g = 0.71$, 95% CI [0.57, 0.85], $Z = 10.0$, $p < 0.001$; B4-B1, $g = 0.78$, 95% CI [0.63, 0.94], $Z = 9.98$, $p < 0.001$; B5-B1, $g = 0.78$, 95% CI [0.63, 0.94], $Z = 10.0$, $p < 0.001$). These results indicate that variation in the actual correlation value would not substantially modify the reported effect sizes and the overall findings of the current meta-analysis.

Heterogeneity Analysis

Significant heterogeneity was found, suggesting variability between studies and demonstrating the importance of accounting for study-level moderators.

B2-B1, $Q_{(44)} = 89.4$, $p < 0.001$, $I^2 = 50.8$; B3-B1, $Q_{(44)} = 144.1$, $p < 0.001$, $I^2 = 69.5$; B4-B1, $Q_{(44)} = 160.1$, $p < 0.001$, $I^2 = 72.5$; B5-B1, $Q_{(44)} = 89.4$, $p < 0.001$, $I^2 = 50.8$.

Moderation Analysis

None of the moderators described below approached significance, indicating that the heterogeneity between studies is not explained by increased age, differences in years of education or proportion of females.

B2-B1: age, $b = 0.01$, CI 95% [−0.01, 0.03], $Z_{(44)} = 0.82$, $p = 0.410$; years of education, $b = -0.0002$, CI 95% [−0.04, 0.04], $Z_{(35)} = -0.11$, $p = 0.912$; proportion of females, $b = 0.46$, CI 95% [−0.27, 1.19], $Z_{(35)} = 1.22$, $p = 0.221$.

B3-B1: age, $b = -0.004$, CI 95% [−0.03, 0.02], $Z_{(44)} = -0.28$, $p = 0.777$; years of education, $b = 0.0038$, CI 95% [−0.05, 0.05], $Z_{(35)} = 0.15$, $p = 0.880$; proportion of females, $b = 0.37$, CI 95% [−0.60, 0.34], $Z_{(35)} = 0.75$, $p = 0.453$.

B4-B1: age, $b = -0.004$, CI 95% [−0.04, 0.03], $Z_{(44)} = -0.26$, $p = 0.795$; years of education, $b = -0.004$, CI 95% [−0.06, 0.049], $Z_{(35)} = -0.15$, $p = 0.882$; proportion of females, $b = 0.0007$, CI 95% [−1.07, 1.07], $Z_{(35)} = 0.00$, $p = 0.999$.

B5-B1: age, $b = -0.005$, CI 95% [−0.04, 0.03], $Z_{(44)} = -0.29$, $p = 0.770$; years of education, $b = -0.004$, CI 95% [−0.06, 0.05], $Z_{(35)} = -0.16$, $p = 0.873$; proportion of females, $b = 0.0018$, CI 95% [−1.06, 1.06], $Z_{(35)} = 0.00$, $p = 0.997$.

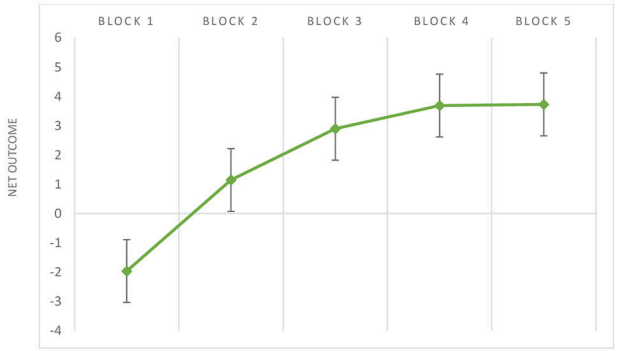


FIGURE 2 | Mean values (and standard errors) of net outcomes (y-axis) considering the performance of older adults across IGT blocks (x-axis).

DISCUSSION

The IGT is one of the most widely used tools to assess decision-making. However, most of the research on IGT and aging has been mainly focused on the performance comparison between older adults and clinical or younger groups. Despite the evidence that older adults make more disadvantageous decisions than younger groups on the IGT (Mata et al., 2011), one question remains unclear: do older adults learn to choose advantageously along the task?

The trend to collapse the choices across blocks to create a summary score—the total net score—restricts the understanding of learning processes that may take place during the IGT. In fact, and to our knowledge, this study is the first to consider within-subject methods when meta-analyzing the IGT performance in different blocks. This method is a relevant contribution to the research field of decision-making in older-adults, as performance in IGT may be ruled by two distinct types of decision-making—under uncertainty and under risk—and only the first is found to be impaired in older adults (Mata et al., 2011). Therefore, it is critical to understand if older adults' decisions remain ruled by uncertainty or, in turn, older adults are capable to learn from experience and move to decisions based on known outcomes.

During the first blocks, the decision-making on the IGT is expected to be driven by affective cues. This is an exploratory stage of learning, as participants have not yet deciphered the contingencies of the decks, and decision-making is made under uncertainty (Brand et al., 2007). Confirming the exploratory process of learning, a negative block net score on block 1 stands out in the older group. Right after block 1, a significant reversal learning effect was found. However, the effect size in relation to the difference between block 2 and block 1 was only small-to-medium in magnitude ($g = 0.48$), demonstrating that a significant improvement in performances is not a robust finding across studies.

From trial 50 onward, choices are expected to be more adaptive and driven by the acquired knowledge (Bechara, 2007). The decision is now expected to be made under risk, as the contingencies of the task are expected to be learned (Brand et al., 2007). The effect sizes from block 3 to block 5 became medium-to-large in magnitude ($g = 0.70$ to 0.78), which is in line with the literature defining the trial 50 as the starting point to develop adaptive choices on the IGT (Bechara, 2007).

The main findings provide evidence that older adults exhibit an advantageous pattern of performance during the IGT. The robust reversal learning effect evidenced in block 3 suggests that the shift from ambiguity to risk seems to occur in this block, and importantly, around the trial proposed by Bechara and colleagues (Bechara, 2007). The hypothesis that older adults, compared to younger groups, tend to choose immediately attractive options on IGT that lead to higher monetary losses along the task (Mata

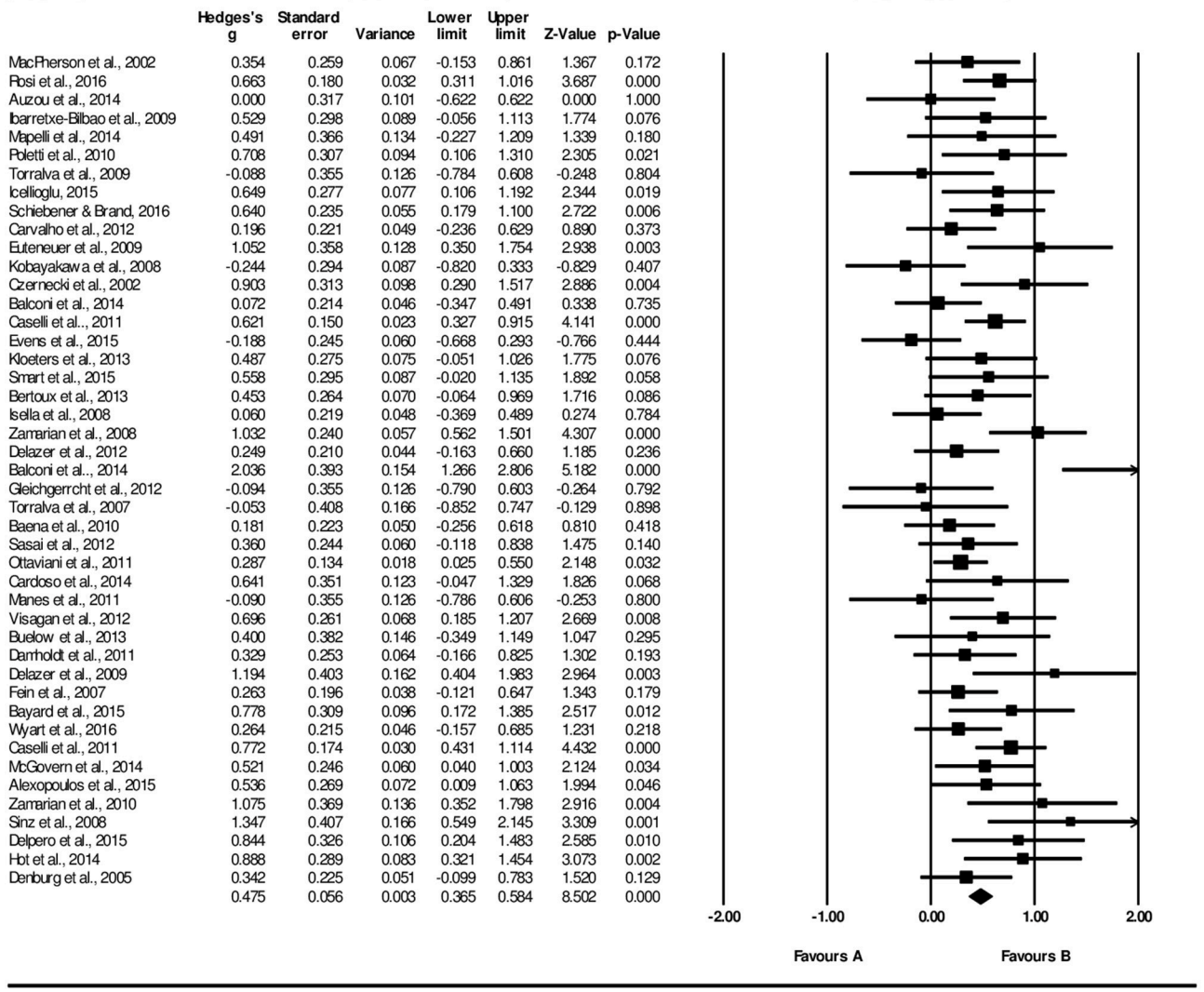


FIGURE 3 | Forest plot of older performance in the initial IGT block (B2-B1).

et al., 2011) does not necessarily mean that older adults are not capable to learn under uncertainty. The within-subjects design of our meta-analysis highlights that learning processes under uncertainty are not entirely compromised with increased age. Moreover, our analyses illustrate how collapsing the choices in a total net score might hide the reversal learning effect across the blocks, masking the older adults' ability to learn under uncertainty.

From our results, older adults seem to be able to use the salient affective stimuli and then integrate these somatic markers (Damasio, 1994) in memory and rational analytical systems, albeit in a less effective way than younger adults (Frank and Kong, 2008; Hämmerer et al., 2011; Mata et al., 2011).

The reviewed studies indicate that older adults show a positive net outcome while performing the task, which means that they finish the task with an adaptive pattern of decision-making, by choosing the advantageous decks more frequently. Only 8 of the 43 studies reported a negative performance in older adults. Despite the positive outcome evidenced by older adults, it should be acknowledged that a net score of $\geq$10 is the cut-off index that describe performances that are not within the range of vmPFC patients (Bechara and Damasio, 2002; Bechara et al., 2002). The total net normative value of older adults is below 10 by Bechara's criterion (Bechara et al., 2002), which would suggest impaired performance and a "myopia for the future" in older groups. The variance around the mean must be taken, however, into account ($\pm$25.9), as well as individual differences.

Direct evidence from the 8 studies reporting negative net scores (Bakos et al., 2008; Wagner et al., 2009; Ottaviani

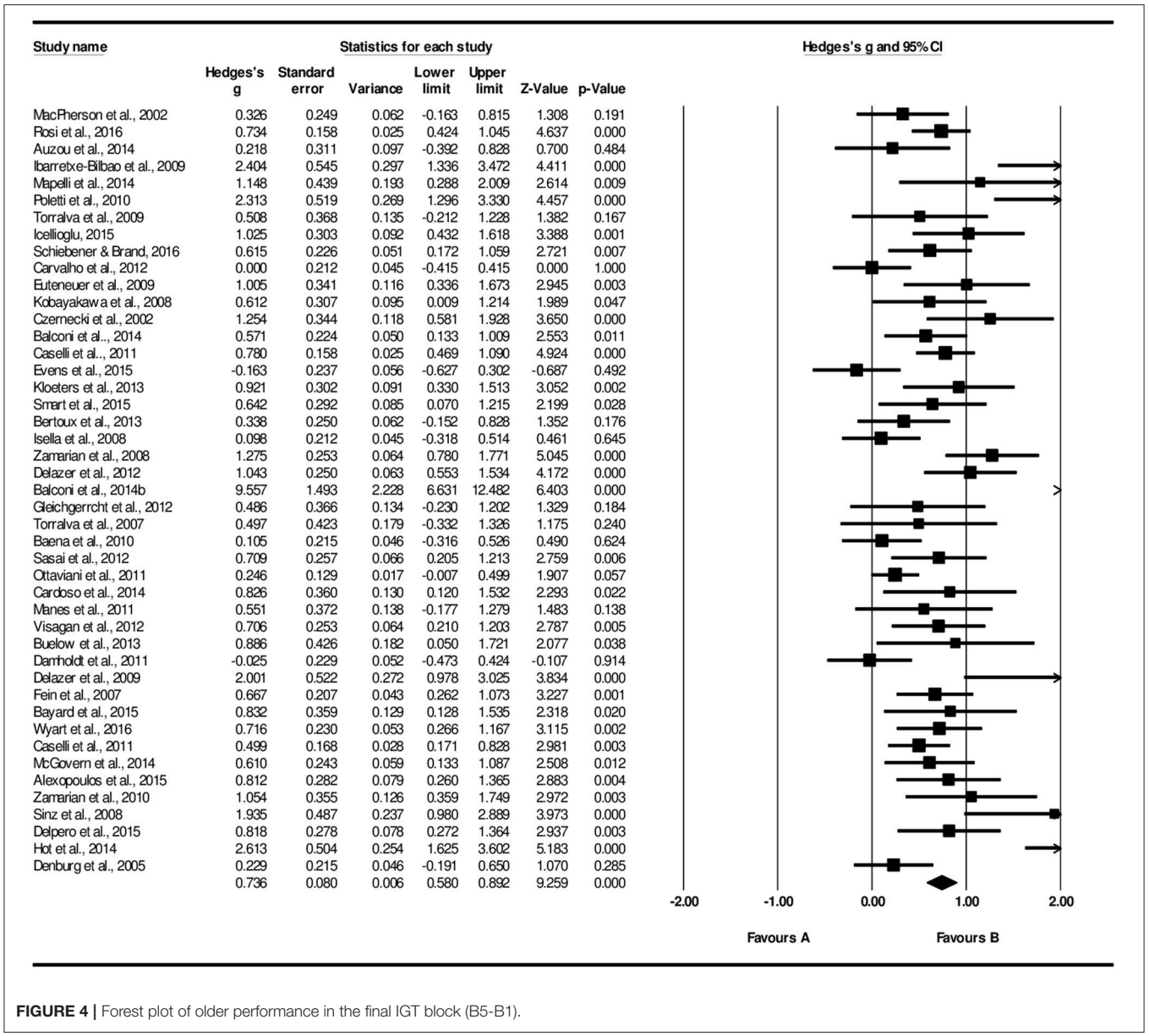


FIGURE 4 | Forest plot of older performance in the final IGT block (B5-B1).

and Vandone, 2011; Bertoux et al., 2013; Auzou et al., 2014; Evens et al., 2015; Schiebener and Brand, 2016) may help to identify relevant individual differences implicated in IGT performance, since the moderators systematically reported in the literature failed to achieve significance in our meta-analysis. This comprehensive analysis is limited, however, by the focus of the included studies on group differences. The focus of the majority of the studies reporting negative net outcomes is redirected to variables that explain impaired performance in clinical groups as opposed to a comprehensive interpretation of the performance of healthy older groups. Nevertheless, Bakos et al. (2008) observed that the oldest old group exhibited poor decision-making in IGT compared to the younger elderly group, despite similar performance in selective attention, short-term memory, and working memory. This finding would suggest that increasing age may compromise adaptive performance in IGT, but our meta-regression showed that age did not moderate the findings. In turn, Schiebener and Brand (2016) included an age range of 18–86 years. Remarkably, age-related variance on IGT performance occurred only in the last 60 trials and in a task with explicit instructions, that is, when decisions are expected to be conducted under risk (Schiebener and Brand, 2016). In the first 40 trials—decision-making under uncertainty—the association between increasing age and less advantageous decision-making was small. This main finding is in line with the theoretical background of the current meta-analysis, highlighting that economic decision-making in later life shows specific dynamics.

Under uncertainty, the amygdala is a critical brain area to trigger affective cues and respond to primary inducers (Damasio,

1994; Hsu et al., 2005) and, interestingly, age does not seem to significantly affect this structure (Mather et al., 2004). This may explain why older adults are capable of deciding advantageously under uncertainty. The difficulty in achieving a performance similar to the younger group, as previously documented in Mata et al.'s (2011) meta-analysis, may be explained by a functional deficit in cognitive functioning. Schiebener and Brand (2016) reported that, after controlling for the effects of cognitive abilities, no age-related variance in decision-making in the IGT remained. This result suggests that age-related changes in EF and reasoning may explain individual differences in IGT performance.

The differences between older and younger adults may be further explained by the difficulty in the older group to persecute the option more likely to be rewarded when differences in reward likelihood are small (Hämmerer et al., 2011). Steingroever et al. (2013) argued that 3 of 4 decks seem to present too similar outcomes. Reduced and similar FRN amplitude in the processing of gains and losses in older adults was previously documented (Hämmerer et al., 2011).

Age-related effects on risky decision-making extends beyond cognition and is also linked to individual differences in personality. Denburg et al. (2009) found that high levels of trait neuroticism in older adults (i.e., proneness to experience negative affective states such as fear, anxiety, sadness, guilt, and anger) is associated with impaired decision-making performance. Importantly, younger adults with high trait-anxiety (Suhr and Tsanadis, 2007) and negative affect (Miu et al., 2008) also show impaired decision-making under uncertainty, despite the increased and potentially adaptive anticipatory somatic signals associated with high trait-anxiety (Suhr and Tsanadis, 2007). The main findings suggest that affect and personality are critical mechanisms to extend our knowledge on older adults' performance in the IGT and, therefore, studies should include these variables.

From the empirical evidence accumulated along the years, the normative total net score for older groups is of 7.55 (±25.9). The lack of age moderation effects suggests that the proposed normative score is representative of the 55–79 age range. The calculation of normative scores, even limited to a statistical criterion, constitute a group reference to compare individual performances of IGT in healthy older adults. Future studies may cluster impaired and unimpaired performances from *Z* scores. The Z scores represent the number of standard deviations below or above the mean considering the individual total net score. Moreover, performance may be analyzed depending on whether the overall score is significantly different from the normative pooled mean, in a negative or positive direction, using the binomial test (Siegel, 1956; Damasio, 1994; Denburg et al., 2005). Under the assumption that a total net score of zero reflects equal probability to choose advantageous and disadvantageous decks, impaired performance is significantly different from zero in a negative direction, while unimpaired performance differs significantly in a positive direction (Denburg et al., 2005). The participants whose total score is not statistically significant from zero in either direction may be included in the borderline group (Denburg et al., 2005).

In sum, our results contradict the assumption (Kovalchik and Allman, 2006) that older adults engage in a random selection strategy, since older adults tend to evidence a pattern of advantageous decision-making. Furthermore, our meta-analysis point that the performance within-study variability reported by Steingroever et al. (2013) contrasts with a robust effect size between-studies. Steingroever et al. (2013) proposed that 100 trials were not sufficient to learn to discriminate safe from risky options and, subsequently, the switch behavior from exploration to exploitation would not occur. From our data, older adults seem to first explore the different decks, as evidenced in negative net outcomes in block 1, and then exploit the most profitable options, culminating in positive net outcomes from block 2 onward. The reversal learning effect is consistently found around block 3. These results are line with Bechara's (Bechara et al., 1994; Bechara, 2007) assumption that after 50 selections participants tend to choose the long-term attractive decks. In the later blocks of IGT, older adults decide in some extent toward less risky choices, suggesting that older adults do not remain unconditionally under uncertainty. In turn, the selection strategies seem to be guided in some way by explicit rules acquired in the course of the task.

From our findings, we propose that decision-making on IGT and aging moves toward uncertainty—where the outcomes are unknown—to risk—where the outcomes were learned and may be used to guide adaptive economic decisions. Differences between younger and older groups found in previous studies may be explained by in a great extent by age-related changes in brain areas associated with cold EF and not, necessarily, with impaired reversal learning.

Limitations

This meta-analysis has some limitations that must be taken into account when interpreting the results. Despite the efforts to include gray literature, publication bias was found in the current systematic search, indicating a possible overestimation of the results.

Correlation coefficients between blocks were further imputed since the values means and standard deviations of block net scores were not reported on the original statistical analysis. Importantly, the sensitivity analysis did not alter the overall findings of the meta-analysis. We strongly recommend authors using the IGT in their research to report correlation values between blocks, as well as all the variables systematically identified in the literature thought to modulate IGT performance.

In the current meta-analysis, the included studies were too heterogeneous, but the moderators with satisfactory data points were non-significant. Since the moderators explaining the heterogeneity remain unknown, the use of the normative data and the generalization of the findings may be compromised. A detailed description of variables relevant to assess IGT performance would allow to explore systematically not only the variables accounting for the heterogeneity between-studies, but also to explain the idiosyncrasies on performances evidenced by Steingroever et al. (2013).

Future Directions

Steingroever et al. (2013) proposed that the frequency of losses is an important variable to explain performance on the IGT, given that participants seem to prefer the decks with infrequent losses. Our meta-analysis did not allow to test the trial-to-trial behavior adjustment after losses and gains, again due to the lack of available information. The analysis of gains and losses ratio in function of decks selection is of high relevance to increased caution after losses (Rolison et al., 2013, 2016). This would also help to clarify the pattern of strategies of older adults that may be associated with reduced total net scores.

The existing meta-analyses should also be extended to explore the reversal learning effect in younger (children, adolescents, and younger adults) and clinical populations. Although older adults show a robust learning effect on the expected block, in light of a previous meta-analysis (Mata et al., 2011) it would be important to explore if the elderly need more trials to overcome the initial uncertainty than younger adults. Since a reversal learning effect is observed, we would expect that once under risk (i.e., when the task contingencies were learned) an equivalent performance would be achieved. However, the reversal learning effect may be faster in younger groups, giving them an advantage to reach a more positive total net outcome on the IGT.

Research focused on personality and the IGT also has to be extended to older groups, as it is likely that individual differences modulate age-related changes in IGT performance.

AUTHOR CONTRIBUTIONS

Conceptualization and paper review: RP, AG, CF, FB, FF, and JM. Data codification: RP and AG. Data analysis: RP and FF. Paper preparation: RP. Supervision: CF, FB, FF, and JM.

FUNDING

This research was supported by a Grant ("The Aging Social Brain") from the Fundação BIAL. CF is supported by a doctoral fellowship from the Fundação para a Ciência e a Tecnologia (Grant number: SFRH/BD/112101/2015).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01785/full#supplementary-material

REFERENCES

Alexopoulos, G. S., Manning, K., Kanellopoulos, D., McGovern, A., Seirup, J. K., Banerjee, S., et al. (2015). Cognitive control, reward-related decision making and outcomes of late-life depression treated with an antidepressant. *Psychol. Med*. 45, 3111–3120. doi: 10.1017/S0033291715001075

Areias, G., Paixão, R., and Figueira, A. P. C. (2013). The Iowa gambling task: a critical revision. *Psicol. Teoria e Pesquisa*. 29, 201–210. doi: 10.1590/S0102-37722013000200009

Auzou, N., Foubert-Samier, A., Dupouy, S., and Meissner, W. G. (2014). Facial emotion recognition is inversely correlated with tremor severity in essential tremor. *J. Neural. Transm*. 121, 347–351. doi: 10.1007/s00702-013-1110-1

Baena, E., Allen, P. A., Kaut, K. P., and Hall, R. J. (2010). On age differences in prefrontal function: the importance of emotional/cognitive integration. *Neuropsychologia* 48, 319–333. doi: 10.1016/j.neuropsychologia.2009.09.021

Bakos, D. S., Couto, M., Melo, W. V., Parente, M., Koller, S. H., and Bizarro, L. (2008). Executive functions in the young elderly and oldest old: a preliminary comparison emphasizing decision making. *Psychol. Neurosci*. 1, 183–189. doi: 10.3922/j.psns.2008.2.011

Balconi, M., Finocchiaro, R., and Campanella, S. (2014b). Reward sensitivity, decisional bias, and metacognitive deficits in cocaine drug addiction. *J. Addict. Med*. 8, 399–406. doi: 10.1097/ADM.0000000000000065

Balconi, M., Finocchiaro, R., and Canavesio, Y. (2014a). Reward-system effect (BAS rating), left hemispheric "unbalance" (alpha band oscillations) and decisional impairments in drug addiction. *Addict. Behav*. 39, 1026–1032. doi: 10.1016/j.addbeh.2014.02.007

Bayard, S., Jacus, J. P., Raffard, S., and Gély-Nargeot, M. C. (2015). Conscious Knowledge and Decision Making Under Ambiguity in Mild Cognitive Impairment and Alzheimer Disease. *Alzheimer Dis. Assoc. Disord*. 29, 357–359. doi: 10.1097/WAD.0000000000000061

Bechara, A. (2007). *Iowa Gambling Task (IGT) Professional Manual*. Lutz, FL: Psychological Assessment Resources.

Bechara, A., and Damasio, A. R. (2005). The somatic marker hypothesis: a neural theory of economic decision. *Games Econ. Behav*. 52, 336–372. doi: 10.1016/j.geb.2004.06.010

Bechara, A., Damasio, A. R., Damasio, H., and Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* 50, 7–15. doi: 10.1016/0010-0277(94)90018-3

Bechara, A., and Damasio, H. (2002). Decision-making and addiction (part I): impaired activation of somatic states in substance dependent individuals when pondering decisions with negative future consequences. *Neuropsychologia* 40, 1675–1689. doi: 10.1016/S0028-3932(02)00015-5

Bechara, A., Damasio, H., and Damasio, A. R. (2003). Role of the amygdala in decision-making. *Ann. N.Y. Acad. Sci*. 985, 356–369. doi: 10.1111/j.1749-6632.2003.tb07094.x

Bechara, A., Dolan, S., and Hindes, A. (2002). Decision-making and addiction (part II): myopia for the future or hypersensitivity to reward? *Neuropsychologia* 40, 1690–1705. doi: 10.1016/S0028-3932(02)00016-7

Bechara, A., Tranel, D., and Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain* 123, 2189–2202. doi: 10.1093/brain/123.11.2189

Bertoux, M., Funkiewiez, A., O'Callaghan, C., Dubois, B., and Hornberger, M. (2013). Sensitivity and specificity of ventromedial prefrontal cortex tests in behavioral variant frontotemporal dementia. *Alzheimers. Dement*. 9, S84–S94. doi: 10.1016/j.jalz.2012.09.010

Best, J. R., Miller, P. H., and Jones, L. L. (2009). Executive functions after age 5: changes and correlates. *Dev Ver*. 29, 180–200. doi: 10.1016/j.dr.2009.05.002

Brand, M., Recknor, E. C., Grabenhorst, F., and Bechara, A. (2007). Decisions under ambiguity and decisions under risk: correlations with executive functions and comparisons of two different gambling tasks with implicit and explicit rules. *J. Clin. Exp. Neuropsychol*. 29, 86–99. doi: 10.1080/13803390500507196

Brevers, D., Bechara, A., Cleeremans, A., and Noël, X. (2013). Iowa Gambling Task (IGT): twenty years after–gambling disorder and IGT. *Front. Psychol*. 4:665. doi: 10.3389/fpsyg.2013.00665

Buelow, M. T., Frakey, L. L., Grace, J., and Friedman, J. H. (2014). The contribution of apathy and increased learning trials to risky decision-making in Parkinson's disease. *Arch. Clin. Neuropsychol*. 29, 100–109. doi: 10.1093/arclin/act065

Cardoso, C. D. O., Branco, L. D., Cotrena, C., Kristensen, C. H., Schneider-Bakos, D. D. G., and Fonseca, R. P. (2014). The impact of frontal and cerebellar lesions on decision making: evidence from the Iowa Gambling Task. *Front. Neurosci*. 8:61. doi: 10.3389/fnins.2014.00061

Carvalho, J. C. N., de Oliveira Cardoso, C., Shneider-Bakos, D., Kristensen, C. H., and Fonseca, R. P. (2012). The effect of age on decision making according to the Iowa gambling task. *Span. J. Psychol.* 15, 480–486. doi: 10.5209/rev_SJOP.2012.v15.n2.38858

Caselli, R. J., Dueck, A. C., Locke, D. E. C., Hoffman-Snyder, C. R., Woodruff, B. K., Rapcsak, S. Z., et al. (2011). Longitudinal modeling of frontal cognition in APOE ε4 homozygotes, heterozygotes, and noncarriers. *Neurology* 76, 1383–1388. doi: 10.1212/WNL.0b013e3182167147

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* 10, 101–129. doi: 10.2307/3001666

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn*. New York, NY: Academic Press.

Czernecki, V., Pillon, B., Houeto, J. L., Pochon, J. B., Levy, R., and Dubois, B. (2002). Motivation, reward, and Parkinson's disease: influence of dopatherapy. *Neuropsychologia* 40, 2257–2267. doi: 10.1016/S0028-3932(02)00108-2

Damasio, A. R. (1994). Descartes' Error: Emotion. *Rationality and the Human Brain*. New York, NY: Putnam Grosset Books.

Damholdt, M. F., Borghammer, P., Larsen, L., and Østergaard, K. (2011). Odor identification deficits identify Parkinson's disease patients with poor cognitive performance. *Mov. Disord.* 26, 2045–2050. doi: 10.1002/mds.23782

Defoe, I. N., Dubas, J. S., Figner, B., and van Aken, M. A. (2015). A meta-analysis on age differences in risky decision making: adolescents versus children and adults. *Psychol. Bull.* 141, 48–84. doi: 10.1037/a0038088

Delazer, M., Högl, B., Zamarian, L., Wenter, J., Ehrmann, L., Gschliesser, V., et al. (2012). Decision making and executive functions in REM sleep behavior disorder. *Sleep* 35, 667–673. doi: 10.5665/sleep.1828

Delazer, M., Sinz, H., Zamarian, L., Stockner, H., Seppi, K., Wenning, G. K., et al. (2009). Decision making under risk and under ambiguity in Parkinson's disease. *Neuropsychologia* 47, 1901–1908. doi: 10.1016/j.neuropsychologia.2009.02.034

Delpero, C., Mioni, G., Rubio, J. L., Juárez Ramos, V., Gómez Milán, E., and Stablum, F. (2015). Decision-making and feedback sensitivity: a comparison between older and younger adults. *J. Cogn. Psychol.* 27, 882–897. doi: 10.1080/20445911.2015.1036759

Denburg, N. L., Tranel, D., and Bechara, A. (2005). The ability to decide advantageously declines prematurely in some normal older persons. *Neuropsychologia* 43, 1099–1106. doi: 10.1016/j.neuropsychologia.2004.09.012

Denburg, N. L., Weller, J. A., Yamada, T. H., Shivapour, D. M., Kaup, A. R., LaLoggia, A., et al. (2009). Poor decision making among older adults is related to elevated levels of neuroticism. *Ann. Behav. Med.* 37, 164–172. doi: 10.1007/s12160-009-9094-7

Dickersin, K. (2015). "Publication bias: recognizing the problem, understanding its origins and scope, and preventing harm," in *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments,* eds H. R. Rothstein, A. J. Sutton, and M. Borenstein (Chichester: John Wiley & Sons, Ltd.), 11–33.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., and Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychol. Methods* 1, 170–177. doi: 10.1037/1082-989X.1.2.170

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ* 315, 629–634. doi: 10.1136/bmj.315.7109.629

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Q. J. Econ.* 75, 643–669. doi: 10.2307/1884324

Euteneuer, F., Schaefer, F., Stuermer, R., Boucsein, W., Timmermann, L., Barbe, M. T., et al. (2009). Dissociation of decision-making under ambiguity and decision-making under risk in patients with Parkinson's disease: a neuropsychological and psychophysiological study. *Neuropsychologia* 47, 2882–2890. doi: 10.1016/j.neuropsychologia.2009.06.014

Evans-Roberts, C. E., and Turnbull, O. H. (2010). Remembering relationships: preserved emotion-based learning in Alzheimer's disease. *Exp. Aging Res.* 37, 1–16. doi: 10.1080/0361073X.2011.536750

Evens, R., Stankevich, Y., Dshemuchadse, M., Storch, A., Wolz, M., Reichmann, H., et al. (2015). The impact of Parkinson's disease and subthalamic deep brain stimulation on reward processing. *Neuropsychologia* 75, 11–19. doi: 10.1016/j.neuropsychologia.2015.05.005

Fein, G., McGillivray, S., and Finn, P. (2007). Older adults make less advantageous decisions than younger adults: cognitive and psychological correlates. *J. Int. Neuropsychol. Soc.* 13, 480–489. doi: 10.1017/S135561770707052X

Fernie, G., and Tunney, R. J. (2006). Some decks are better than others: the effect of reinforcer type and task instructions on learning in the Iowa Gambling Task. *Brain Cogn.* 60, 94–102. doi: 10.1016/j.bandc.2005.09.011

Ferreira-Santos, F. (in press). Meta-analysis of correlated designs (repeated measures) in the context of neural responses to facial expressions of emotion. *SAGE Res. Methods Cases: Psychol.*

Field, A. P. (2001). Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. *Psychol. Methods* 6, 161–180. doi: 10.1037/1082-989X.6.2.161

Frank, M. J., and Kong, L. (2008). Learning to avoid in older age. *Psychol. Aging.* 23, 392–398. doi: 10.1037/0882-7974.23.2.392

Gleichgerrcht, E., Torralva, T., Roca, M., Szenkman, D., Ibanez, A., Richly, P., et al. (2012). Decision making cognition in primary progressive aphasia. *Behav. Neurol.* 25, 45–52. doi: 10.1155/2012/606285

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N., Fristona, K. J., and Frackowiak, R. S., (2001). A Voxel-Based Morphometric Study of Ageing in 465 normal adult human brains. *Neuroimage* 14, 21–36. doi: 10.1006/nimg.2001.0786

Hämmerer, D., Li, S. C., Müller, V., and Lindenberger, U. (2011). Life span differences in electrophysiological correlates of monitoring gains and losses during probabilistic reinforcement learning. *J. Cogn. Neurosci.* 23, 579–592. doi: 10.1162/jocn.2010.2147

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *J. Educ. Behav. Stat.* 6, 107–128. doi: 10.3102/10769986006002107

Higgins, J. P., and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions,* Vol. 4. Chichester: John Wiley & Sons.

Hot, P., Ramdeen, K. T., Borg, C., Bollon, T., and Couturier, P. (2014). Impaired decision making in alzheimer's disease a deficit of cognitive strategy selection? *Clin. Psychol. Sci.* 2, 328–335. doi: 10.1177/2167702613504094

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310, 1680–1683. doi: 10.1126/science.1115327

Huang, Y. H., Wood, S., Berger, D. E., and Hanoch, Y. (2015). Age differences in experiential and deliberative processes in unambiguous and ambiguous decision making. *Psychol. Aging* 30, 675–687. doi: 10.1037/pag0000038

Ibarretxe-Bilbao, N., Junque, C., Tolosa, E., Marti, M. J., Valldeoriola, F., Bargallo, N., et al. (2009). Neuroanatomical correlates of impaired decision-making and facial emotion recognition in early Parkinson's disease. *Eur. J. Neurosci.* 30, 1162–1171. doi: 10.1111/j.1460-9568.2009.06892.x.

Icellioglu, S. (2015). Iowa Gambling Test: normative data and correlation with executive functions. *Dusunen. Adam.* 28, 222–230. doi: 10.5350/DAJPN2015280305

Isella, V., Mapelli, C., Morielli, N., Pelati, O., Franceschi, M., and Appollonio, I. M. (2008). Age-related quantitative and qualitative changes in decision making ability. *Behav. Neurol.* 19, 59–56. doi: 10.1155/2008/893727

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291. doi: 10.2307/1914185

Kloeters, S., Bertoux, M., O'Callaghan, C., Hodges, J. R., and Hornberger, M. (2013). Money for nothing—Atrophy correlates of gambling decision making in behavioural variant frontotemporal dementia and Alzheimer's disease. *Neuroimage Clin.* 2, 263–272. doi: 10.1016/j.nicl.2013.01.011

Kobayakawa, M., Koyama, S., Mimura, M., and Kawamura, M. (2008). Decision making in Parkinson's disease: analysis of behavioral and physiological patterns in the Iowa gambling task. *Mov. Disord.* 23, 547–552. doi: 10.1002/mds.21865

Kovalchik, S., and Allman, J. (2006). Measuring reversal learning: introducing the variable iowa gambling task in a study of young and old normals. *Cogn. Emot.* 20, 714–728. doi: 10.1080/02699930500371166

Lawrence, N. S., Jollant, F., O'Daly, O., Zelaya, F., and Phillips, M. L. (2009). Distinct roles of prefrontal cortical subregions in the Iowa Gambling Task. *Cereb. Cortex.* 19, 1134–1143. doi: 10.1093/cercor/bhn154

Li, X., Lu, Z. L., D'Argembeau, A., Ng, M., and Bechara, A. (2010). The Iowa gambling task in fMRI images. *Hum. Brain Mapp.* 31, 410–423. doi: 10.1002/hbm.20875.

MacPherson, S. E., Phillips, L. H., and Della Sala, S. (2002). Age, executive function and social decision making: a dorsolateral prefrontal theory of cognitive aging. *Psychol. Aging* 17, 598–609. doi: 10.1037/0882-7974.17.4.598

Manes, F., Torralva, T., Ibáñez, A., Roca, M., Bekinschtein, T., and Gleichgerrcht, E. (2011). Decision-making in frontotemporal dementia: clinical, theoretical and legal implications. *Dement. Geriatr. Cogn. Disord.* 32, 11–17. doi: 10.1159/000329912.

Mapelli, D., Di Rosa, E., Cavalletti, M., Schiff, S., and Tamburin, S. (2014). Decision and dopaminergic system: an ERPs study of Iowa gambling task in Parkinson's disease. *Front. Psychol.* 5:684. doi: 10.3389/fpsyg.2014.00684

Mata, R., Josef, A. K., Samanez-Larkin, G. R., and Hertwig, R. (2011). Age differences in risky choice: a meta-analysis. *Ann. N.Y. Acad. Sci.* 1235, 18–29. doi: 10.1111/j.1749-6632.2011.06200.x

Mather, M., Canli, T., English, T., Whitfield, S., Wais, P., Ochsner, K., et al. (2004). Amygdala responses to emotionally valenced stimuli in older and younger adults. *Psychol. Sci.* 15, 259–263. doi: 10.1111/j.0956-7976.2004.00662.x

McGovern, A. R., Alexopoulos, G. S., Yuen, G. S., Morimoto, S. S., and Gunning-Dixon, F. M. (2014). Reward-related decision making in older adults: relationship to clinical presentation of depression. *Int. J. Geriatr. Psychiatry*. 29, 1125–1131. doi: 10.1002/gps.4200

Miu, A. C., Heilman, R. M., and Houser, D. (2008). Anxiety impairs decision-making: psychophysiological evidence from an Iowa Gambling Task. *Biol. Psychol.* 77, 353–358. doi: 10.1016/j.biopsycho.2007.11.010.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cogn. Psychol.* 41, 49–100. doi: 10.1006/cogp.1999.0734

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS. Med*. 6:E1000097. doi: 10.1371/journal.pmed.1000097

Morris, S. B., and DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol. Methods* 7, 105–125. doi: 10.1037/1082-989X.7.1.105

Northoff, G., Grimm, S., Boeker, H., Schmidt, C., Bermpohl, F., Heinzel, A., et al. (2006). Affective judgment and beneficial decision making: ventromedial prefrontal activity correlates with performance in the Iowa Gambling Task. *Hum. Brain Mapp.* 27, 572–587. doi: 10.1002/hbm.20202

Ottaviani, C., and Vandone, D. (2011). Impulsivity and household indebtedness: evidence from real life. *J. Econ. Psychol.* 32, 754–761. doi: 10.1016/j.joep.2011.05.002

Pagonabarraga, J., García-Sánchez, C., Llebaria, G., Pascual-Sedano, B., Gironell, A., and Kulisevsky, J. (2007). Controlled study of decision-making and cognitive impairment in Parkinson's disease. *Mov. Disord.* 22, 1430–1435. doi: 10.1002/mds.21457

Perretta, J. G., Pari, G., and Beninger, R. J. (2005). Effects of Parkinson disease on two putative nondeclarative learning tasks: probabilistic classification and gambling. *Cogn. Behav. Neurol.* 18, 185–192. doi: 10.1097/01.wnn.0000187939.81541.1d

Poletti, M., Frosini, D., Lucetti, C., Dotto, P. D., Ceravolo, R., and Bonuccelli, U. (2010). Decision making in de novo Parkinson's disease. *Mov. Disord.* 25, 1432–1436. doi: 10.1002/mds.23098

Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., et al. (2005). Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cereb Cortex.* 15, 1676–1689. doi: 10.1093/cercor/bhi044

Resnick, S. M., Lamar, M., and Driscoll, I. (2007). Vulnerability of the orbitofrontal cortex to age-associated structural and functional brain changes. *Ann. N.Y. Acad. Sci.* 1121, 562–575. doi: 10.1196/annals.1401.027

Rolison, J. J., Hanoch, Y., Wood, S., and Liu, P. J. (2013). Risk-taking differences across the adult life span: a question of age and domain. *J. Gerontol. B. Psychol. Sci. Soc. Sci.* 69, 870–880. doi: 10.1093/geronb/gbt081

Rolison, J. J., Wood, S., and Hanoch, Y. (2016). Age and adaptation: stronger decision updating about real world risks in older age. *Risk Anal.* 37, 1632–1643. doi: 10.1111/risa.12710

Rosi, A., Cavallini, E., Gamboz, N., and Russo, R. (2016). On the generality of the effect of experiencing prior gains and losses on the Iowa Gambling Task: a study on young and old adults. *Judgm. Decis. Mak.* 11:185. Available online at: http://journal.sjdm.org/15/151216/jdm151216.html

Sasai, T., Miyamoto, T., Miyamoto, M., Iwanami, M., Abe, T., Matsuura, M., et al. (2012). Impaired decision-making in idiopathic REM sleep behavior disorder. *Sleep Med.* 13, 301–306. doi: 10.1016/j.sleep.2011.09.012

Schiebener, J., and Brand, M. (2016). Age-related variance in decisions under ambiguity is explained by changes in reasoning, executive functions, and decision-making under risk. *Cogn. Emot.* 22, 1–9. doi: 10.1080/02699931.2016.1159944

Schmidt, F. L., Oh, I. S., and Hayes, T. L. (2009). Fixed- versus random-effects models in metaanalysis: Model properties and an empirical comparison of differences in results. *Br. J. Math. Stat. Psychol.* 62, 97–128. doi: 10.1348/000711007X255327

Schneider, D., and Parente, M. A. (2006). O desempenho de adultos jovens e idosos na Iowa Gambling Task (IGT): um estudo sobre a tomada de decisão. *Psicol. Reflexão e Crítica* 19, 442–450. doi: 10.1590/S0102-79722006000300013

Séguin, J. R., Arseneault, L., and Tremblay, R. E. (2007). The contribution of "cool" and "hot" components of decision-making in adolescence: implications for developmental psychopathology. *Cogn. Dev.* 22, 530–543. doi: 10.1016/j.cogdev.2007.08.006

Shivapour, S. K., Nguyen, C. M., Cole, C. A., and Denburg, N. L. (2012). Effects of age, sex, and neuropsychological performance on financial decision-making. *Front. Neurosci.* 6:82. doi: 10.3389/fnins.2012.00082

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. New York, NY: McGraw Hill.

Silver, N. C., and Dunlap, W. P. (1987). Averaging correlation coefficients: should Fisher's z transformation be used? *J. Appl. Psychol.* 72, 146–148.

Sinz, H., Zamarian, L., Benke, T., Wenning, G. K., and Delazer, M. (2008). Impact of ambiguity and risk on decision making in mild Alzheimer's disease. *Neuropsychologia* 46, 2043–2055. doi: 10.1016/j.neuropsychologia.2008.02.002

Smart, C. M., and Krawitz, A. (2015). The impact of subjective cognitive decline on Iowa Gambling Task performance. *Neuropsychology* 29, 971–987. doi: 10.1037/neu0000204

Sowell, E. R., Peterson, B. S., Thompson, P. M., Welcome, S. E., Henkenius, A. L., and Toga, A. W. (2003). Mapping cortical change across the human life span. *Nat. Neurosci.* 6, 309–315. doi: 10.1038/nn1008

Steingroever, H., Wetzels, R., Horstmann, A., Neumann, J., and Wagenmakers, E. J. (2013). Performance of healthy participants on the Iowa gambling task. *Psychol. Assessment*. 25: 180. doi: 10.1037/a0029929

Suhr, J. A., and Tsanadis, J. (2007). Affect and personality correlates of the Iowa Gambling Task. *Person. Individ. Differ.* 43. 27–36. doi: 10.1016/j.paid.2006.11.004

Torralva, T., Kipps, C. M., Hodges, J. R., Clark, L., Bekinschtein, T., Roca, M., et al. (2007). The relationship between affective decision-making and theory of mind in the frontal variant of fronto-temporal dementia. *Neuropsychologia* 45, 342–349. doi: 10.1016/j.neuropsychologia.2006.05.031

Torralva, T., Roca, M., Gleichgerrcht, E., Bekinschtein, T., and Manes, F. (2009). A neuropsychological battery to detect specific executive and social cognitive impairments in early frontotemporal dementia. *Brain* 132, 1299–1309. doi: 10.1093/brain/awp041

Verdejo-Garcia, A., Bechara, A., Recknor, E., and Perez-Garcia, M. (2006). Decision-making and the iowa gambling task: ecological validity in individuals with substance dependence. *Psychol. Belg.* 46, 55–78. doi: 10.5334/pb-46-1-2-55

Visagan, R., Xiang, A., and Lamar, M. (2012). Comparison of deck-and trial-based approaches to advantageous decision making on the Iowa Gambling Task. *Psychol. Assess.* 24:455. doi: 10.1037/a0025932

Wagner, G. P., Trentini, C. M., and Parente, M. A. D. M.P. (2009). O desempenho de idosos com e sem declínio cognitivo leve nos Testes Wisconsin de Classificação de Cartas e Iowa Gambling Test. *Psico* 40, 220–226.

Walhovd, K. B., Fjell, A. M., Reinvang, I., Lundervold, A., Dale, A. M., Eilertsen, D. E., et al. (2005). Effects of age on volumes of cortex, white matter and subcortical structures. *Neurobiol. Aging.* 26, 1261–1270. doi: 10.1016/j.neurobiolaging.2005.05.020

West, R. (2000). In defense of the frontal lobe hypothesis of cognitive aging. *J. Int. Neuropsych. Soc.* 6, 727–729.

Wyart, M., Jaussent, I., Ritchie, K., Abbar, M., Jollant, F., and Courtet, P. (2016). Iowa Gambling Task performance in elderly persons with a lifetime history of suicidal acts. *Am. J. Geriatr. Psychiatry*. 24, 399–406. doi: 10.1016/j.jagp.2015.12.007

Zamarian, L., Sinz, H., Bonatti, E., Gamboz, N., and Delazer, M. (2008). Normal aging affects decisions under ambiguity, but not decisions under risk. *Neuropsychology* 22, 645–657. doi: 10.1037/0894-4105.22.5.645

Zamarian, L., Weiss, E. M., and Delazer, M. (2011). The impact of mild cognitive impairment on decision making in two gambling tasks. *J. Gerontol. B Psychol. Sci. Soc. Sci*. 66, 23–31. doi: 10.1093/geronb/gbq067

Zelazo, P. D., and Müller, U. (2002). "Executive function in typical and atypical development," in *Handbook of Childhood Cognitive De velopment,* ed U. Goswami, (Oxford: Blackwell), 445–469.

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: 10 August 2018
doi: 10.3389/fpsyg.2018.01461

Cognitive Style and Frame Susceptibility in Decision-Making

David R. Mandel[1,2]* *and Irina V. Kapler*[2]

[1] *Intelligence Group, Intelligence, Influence and Collaboration Section, Defence Research and Development Canada, Toronto, ON, Canada,* [2] *Department of Psychology, York University, Toronto, ON, Canada*

The susceptibility of decision-makers' choices to variations in option framing has been attributed to individual differences in cognitive style. According to this view, individuals who are prone to a more deliberate, or less intuitive, thinking style are less susceptible to framing manipulations. Research findings on the topic, however, have tended to yield small effects, with several studies also being limited in inferential value by methodological drawbacks. We report two experiments that examined the value of several cognitive-style variables, including measures of cognitive reflection, subjective numeracy, actively open-minded thinking, need for cognition, and hemispheric dominance, in predicting participants' frame-consistent choices. Our experiments used an isomorph of the Asian Disease Problem and we manipulated frames between participants. We controlled for participants' sex and age, and we manipulated the order in which choice options were presented to participants. In Experiment 1 (N = 190) using an undergraduate sample and in Experiment 2 (N = 316) using a sample of Amazon Mechanical Turk workers, we found no significant effect of any of the cognitive-style measures taken on predicting frame-consistent choice, regardless of whether we analyzed participants' binary choices or their choices weighted by the extent to which participants preferred their chosen option over the non-chosen option. The sole factor that significantly predicted frame-consistent choice was framing: in both experiments, participants were more likely to make frame-consistent choices when the frame was positive than when it was negative, consistent with the tendency toward risk aversion in the task. The present findings do not support the view that individual differences in people's susceptibility to framing manipulations can be substantially accounted for by individual differences in cognitive style.

Keywords: framing effect, risky choice, Asian disease problem, cognitive style, individual differences

OPEN ACCESS

Edited by:
Bernhard Hommel,
Leiden University, Netherlands

Reviewed by:
Li-Lin Rao,
Institute of Psychology (CAS), China
Anthony John Porcelli,
Marquette University, United States

***Correspondence:**
David R. Mandel
drmandel66@gmail.com

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

Received: *18 March 2018*
Accepted: *25 July 2018*
Published: *10 August 2018*

Citation:
Mandel DR and Kapler IV (2018) Cognitive Style and Frame Susceptibility in Decision-Making. Front. Psychol. 9:1461. doi: 10.3389/fpsyg.2018.01461

INTRODUCTION

Literature on risky choice shows that, in general, people are susceptible to a wide range of framing effects. Such effects signal incoherence in decision-making because they ostensibly violate the description invariance principle, which states that mere re-descriptions of events that do not alter their extension should, likewise, not alter people's choices. The description invariance principle is one of the least controversial coherence principles undergirding rational choice theory, and thus violations of it are regarded as prima facie evidence of irrationality in human decision-making (Arrow, 1982; Tversky and Kahneman, 1986).

The most frequently studied type of framing effect involves re-describing the possible outcomes of two alternative options in terms that are meant to either emphasize gain (positivity)

or loss (negativity). Tversky and Kahneman's (1981) Asian Disease Problem (ADP) provides a seminal demonstration of the manipulation. All participants first read the following:

Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

Participants in the positive-framing condition chose between the following options:

If Program A is adopted, 200 people will be saved.

If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.

Participants in the negative-framing condition instead chose between these options:

If Program C is adopted, 400 people will die.

If Program D is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

Seventy-two percent chose the certain option (A) under positive framing, whereas 78% chose the uncertain option (D) under negative framing.[1]

The framing effect demonstrated using the ADP and problem isomorphs—namely, the tendency to make risk-averse choices given positive frames (option-A choices) and risk-seeking choices given negative frames (option-D choices)—is highly replicable and tends to yield small to moderate effect sizes (for meta-analyses, see Kühberger, 1998; Piñon and Gambara, 2005). Although much of the literature on framing effects and the ADP, in particular, has focused on nomothetic patterns of choice, it is evident that susceptibility to frame-consistent choices exhibits individual differences, which theoretical accounts of framing tend to neglect. For instance, individual differences in the linguistic interpretation of numerical quantifiers in the two options of the ADP influence the proportion of samples showing the standard framing effect. Mandel (2014, Experiment 3) found that the standard framing effect was observed among participants who interpreted the numeric quantifier in the certain option as a lower bound (i.e., meaning *at least*), but it was not observed among participants who interpreted the same quantifiers as meaning an exact value (i.e., either *exactly* 200 will be saved in the positive frame or *exactly* 400 will die in the negative frame). A minority of participants who interpreted the same quantifiers as representing an upper bound (i.e., *at most*. . .) actually showed a reversed framing effect.

By far, however, most attention to individual differences in susceptibility to framing on choice has focused on variations in people's cognitive style. The interest, in part, would seem to stem from a more recent view of such effects as relying on fast and intuitive "System 1" reasoning processes rather than slower, analytic "System 2" reasoning processes (Stanovich and West, 2000; De Martino et al., 2006; Evans, 2008, 2010; Kahneman, 2011). Consistent with this view, some studies find that requiring people to thoughtfully consider problem options (i.e., to shift from System 1 reasoning to the more deliberate and effortful, System 2 reasoning) attenuates framing effects. For example, Takemura (1994) asked participants to either write an open-ended justification for their choices in the ADP or simply choose between the two programs. In the high-elaboration condition, the framing effect was eliminated. Likewise, Almashat et al. (2008) found that deeper deliberation in medical decision-making, achieved by asking participants to list advantages and disadvantages of each treatment option prior to making a choice, reduced framing effects (see also Miller and Fagley, 1991; Sieck and Yates, 1997).

If situational manipulations that affect the degree of deliberateness in decision-making can moderate framing effects, then perhaps individual differences in cognitive-style measures that track deliberateness in thinking might also predict susceptibility to frame-consistent patterns of choice. One such hypothesis is that need for cognition (NFC) moderates framing effects. NFC measures the extent to which individuals enjoy engaging in effortful thinking (Cacioppo et al., 1984). Several studies have explored the relationship between NFC and framing effects. For example, Smith and Levin (1996) found that participants high in NFC showed no framing effect on multiple decision tasks, whereas participants low in NFC showed framing effects (see also Carnevale et al., 2011). However, Simon et al. (2004) in a between-subjects design found that being high in NFC alone was insufficient to eliminate framing effects. Only participants who were high in NFC and high in self-rated math or engaged in deep thinking showed reduced susceptibility to framing effects (also see Covey, 2014). LeBoeuf and Shafir (2003) asked participants to answer a number of framing problems as well as provide justifications for their responses. Although NFC did not moderate the framing effect when participants were exposed to only one frame in a between-subjects design, participants high in NFC were more likely than low-NFC participants to make consistent choices across frames when presented with both frames in a within-subjects design. LeBoeuf and Shafir (2003) posited that NFC increased consistency detection across frames but did not diminish the framing effect when it was impossible for participants to verify their consistency in choice. However, Levin et al. (2002) did not find a relation between NFC and framing even when utilizing within-subjects designs (participants answered both frames options separated by 1 week in time). Likewise, Peters and Levin (2008) and Mandel (2014, Experiment 1) did not find evidence that NFC moderated the framing effect.

A related cognitive-style measure that has been explored as a possible moderator of the framing effect is actively open-minded thinking (AOT). AOT involves a willingness to evaluate evidence that goes against one's beliefs, and openness to considering alternative perspectives (Baron, 1985, 1993; Haran et al., 2013; Baron et al., 2015). AOT and NFC are positively correlated (West et al., 2008, 2012; Haran et al., 2013). AOT has been positively associated with accuracy in forecasting (Tetlock, 2005; Mellers et al., 2015) and other probabilistic judgment tasks

[1]Although, we are aware of Knight's (1964) distinction between risk and uncertainty, we prefer to use the term *uncertain* rather than *risky* for options B and D in the ADP because (a) Mandel (2014) has shown that participants often have imprecise interpretations of ostensibly precise probabilities such as "1/3 chance." Moreover, Tombu and Mandel (2015) have shown that a non-trivial proportion of participants perceive option C as riskier than option D.

(Haran et al., 2013). Actively open-minded thinkers are also less prone to biases, including myside bias (Baron, 2008; West et al., 2008), belief bias, framing, and base-rate neglect (West et al., 2008; Toplak et al., 2011, 2017).

A third cognitive-style measure that might be expected to index the degree of System 2 reasoning is the Cognitive Reflection Test (CRT; Frederick, 2005), which measures individuals' abilities to suppress incorrect intuitive answers and answer correctly. CRT performance is positively related to AOT (e.g., Toplak et al., 2014; Baron et al., 2015) and to performance on risky choice tasks (Cokely and Kelley, 2009). CRT is also negatively related to a wide range of cognitive biases in judgment and decision-making (e.g., Toplak and Stanovich, 2002; Campitelli and Labollita, 2010; Toplak et al., 2011; Baldi et al., 2013). The evidence regarding the relation between framing susceptibility and CRT is mixed, however, with some literature reporting positive relations (Oechssler et al., 2009; Noori, 2016) and other literature reporting no relation (Toplak et al., 2014; Aczel et al., 2015). There is also disagreement about what CRT measures. Sinayev and Peters (2015) posit that CRT mainly taps numeracy skill, whereas Pennycook and Ross (2016) examined evidence showing that CRT predicted a wide range of variables not attributable to numeracy. Szaszi et al. (2017) concluded that both numeracy and cognitive reflection (indicative of System 2 reasoning) account for CRT performance. However, item analysis of the CRT showed that only one of the three items (the bat and ball problem) had faster response time when the answer was the intuitive incorrect response than when it was the correct response (Stupple et al., 2017). Moreover, even in that case, the effect was small. Such findings raise doubt about the extent to which performance on the measure captures "counter-deliberate" cognitive miserliness.

Lipkus and Peters (2009) posited that numeracy has a number of functions, such as facilitating assessment of likelihood and value, improving interpretation and acceptance of numerical data, encouraging information seeking and greater depth of processing. Yet numeracy shows a mixed pattern of evidence in studies on framing. In Peters et al. (2006, Study 1), less numerate participants showed larger framing effects when asked to rate the quality of students' work that was presented either as percent correct (74%) or percent incorrect (26%) on an exam (see also Garcia-Retamero and Galesic, 2010; Peters, 2012). Likewise, Peters and Levin (2008) found that less numerate participants showed larger framing effects on choice than more numerate participants. However, Peters et al. (2011) found that numeracy did not moderate framing effects on a task that assessed medication risk. Whereas numeracy is often *objectively* measured in terms of performance skill (e.g., Lipkus et al., 2001; Weller et al., 2013), it can also be measured using a Subjective Numeracy Scale (SNS), which taps individuals' preferences for processing numbers and graphical information over words (Fagerlin et al., 2007). Individuals lower in subjective numeracy had more negative emotional reactions to numbers and were less motivated and/or confident in numeric tasks (Peters and Bjalkebring, 2015). Compared with objective numeracy scales that require participants to complete calculations, SNS has several advantages (Fagerlin et al., 2007; Kee and Liang, 2015): it takes less time to complete, participants find it more enjoyable, less stressful and less frustrating, and there is direct (Zikmund-Fisher et al., 2007) and indirect (Weller et al., 2013) evidence that SNS is a good approximation of objective numeracy. For instance, moderate correlations (in the 0.4–0.7 range) between objective and subjective numeracy scales also have been reported (Fagerlin et al., 2007; Peters and Bjalkebring, 2015; Gamliel et al., 2016). Moreover, unlike objective numeracy, SNS cannot be exploited by use of a calculator, an issue of concern in online studies such as those we report in this article.

The Present Research

The present research builds on prior work examining how cognitive-style measures relate to susceptibility to framing. Although several studies have examined this issue, there has been little attempt to jointly examine cognitive-style measures as predictors of frame susceptibility. Multi-measure analyses are critical, however, because measures such as NFC, AOT, CRT, and SNS share considerable variance.[2] Several studies also have binned participants into high versus low categories based on median splits. This is usually a poor statistical method of analysis because it adds error, reducing statistical power and increasing the likelihood of Type II errors in many cases (Humphreys and Fleishman, 1974; Cohen, 1983). Moreover, in other cases, dichotomizing continuous variables can lead to spurious statistical significance or Type 1 errors (Maxwell and Delaney, 1993). As well, most studies have taken a rather coarse moderator approach to analysis in which the cognitive-style measure, partitioned into high versus low scores, is crossed with a framing manipulation. Evidence of moderation takes the form of showing that the framing effect is smaller in one group than the other. This method is theoretically imprecise because significant effects of framing do not necessarily conform to theoretical expectation. Consider a hypothetical case of an ADP experiment using a large sample in which 90% of participants in the positive-frame condition choose option A, whereas "only" 75% choose option A in the negative-frame condition. The framing effect may be significant, yet most participants who encountered the negative frame would not have responded as predicted by prospect theory (Kahneman and Tversky, 1979) or other theories of framing, such as the explicated valence account (Tombu and Mandel, 2015; also Wallin et al., 2016) or the fuzzy trace account (Reyna and Brainerd, 1991), all of which make the same predictions for the standard ADP (but which diverge in prediction as well as explanation under other task conditions).

Therefore, in the present research, we examine frame susceptibility in a theoretically precise manner. Participants are said to be frame susceptible or to have made frame-consistent choices if and only if they choose the certain option given a positive frame or else they choose the uncertain option given

[2]West et al. (2012) examined the relations between NFC, AOT, and CRT and a composite measure of responses from judgment and decision-making tasks associated with cognitive biases. These tasks included the ADP, but unfortunately, their analyses do not extend to the specific relations between the cognitive-style measures and the ADP. Their findings, however, indicate that all three measures showed small correlations (˜0.1) with the composite measure of cognitive biases.

a negative frame. This operationalization of frame-consistent choice thus requires the demonstration of what has been referred to in previous literature as bidirectional framing effects (Wang, 1996) or meeting the reference distribution criterion (Mandel, 2001). In two experiments, we test various predictive models of frame susceptibility. To the extent that these predictors significantly predict frame susceptibility, it would therefore provide more compelling evidence of moderation of framing effects than we have seen in previous studies. In addition to examining NFC, AOT, CRT, and SNS, we include participants' sex and age in our analysis. This is important because sex has been shown to moderate framing effects (with females showing stronger framing effects; see Piñon and Gambara, 2005 for a meta-analysis) and there is some evidence for sex differences on measures such as CRT (Frederick, 2005), numeracy (objective: Peters et al., 2011; subjective: Peters and Bjalkebring, 2015) and AOT (Toplak et al., 2017) measures, with males scoring higher than females. In terms of age, there appears to be age-related stability in susceptibility to framing (Mayhorn et al., 2002; Rönnlund et al., 2005; Strough et al., 2011). More generally, description-based (as opposed to experience-based) tasks like the ADP tend to show negligible age effects (Mata et al., 2011). Nevertheless, we statistically control for age in the predictive models tested in the present research as a precautionary measure.

Furthermore, we took methodological precautions that have been overlooked in most previous framing research. We manipulated option order as recommended by Fagley and Miller (1997). In most studies, the certain option is presented first, yet Bar-Hillel et al. (2014) have shown a "reachability bias" in response-option selection favoring the option presented first. Kühberger and Gradl (2013) found that although option order had no effect on choice in the positive-frame condition, a greater proportion of participants in the negative-frame condition chose the uncertain option when it was presented after the certain option than when it was presented initially, contrary to the reachability bias. However, using a substantially larger sample, Schwitzgebel and Cushman (2015) found evidence of an order effect in the ADP consistent with reachability bias. In the negative frame, a significantly greater proportion of participants chose the uncertain option when it was presented first. Likewise, in the positive frame a significantly greater proportion chose the certain option when it was presented first. Thus, the standard presentation (certain option first) might overestimate frame susceptibility in the positive-frame condition and underestimate it in the negative-frame condition.

As noted earlier, Mandel (2014) found that the modal interpretation of numerical quantifiers in the certain option was lower-bounded (i.e., "at least *n* will be saved/will die"). On the basis of those and other findings (e.g., Halberg and Teigen, 2009; Teigen and Nikolaisen, 2009), Mandel (2014) recommended that researchers using the ADP make explicit that the numerical quantifiers are intended to be treated as exact values, in order to increase the likelihood that the assumption of extensional equivalence between reframed options is valid. Accordingly, the present research used an isomorph of the ADP, which stated that the value in the certain option presented was *exactly* that value. Finally, following earlier studies (e.g., Mandel, 2001; Tombu and Mandel, 2015), we examine the effects of our predictors on participants' binary choices as well as on a bi-directional strength of preference measure that weights the chosen option by the degree to which that option is judged preferable to its alternative.

EXPERIMENT 1

Materials and Methods

Prior to the initiation of this research, it was reviewed and approved by the York University's Ethics Review Board and deemed to be in conformance with the standards of the Canadian Tri-Council Research Ethics guidelines. Electronic written informed consent was obtained from all participants, who were also debriefed following the experiment.

We recruited 201 undergraduate students enrolled in a first-year psychology course at York University. Students were awarded course credit for participation. Mean age of the sample was 21.7 years (SD = 4.2), and 74.2% were female. Eleven participants were removed from the analysis because the integrity of the collected data was low. Specifically, participants were removed due to (a) unreasonable time for completion (over 24 h, or under 4 min), (b) below 50% self-reported English proficiency (current range 55–100, M = 92.8, SD = 10.7) or (c) unwarranted age for a university sample (e.g., one participant reported being 10 years old). Data from the remaining 190 participants were analyzed.

Experiment 1, which took 15 min of average to complete (SD = 12.50), was conducted online using the Qualtrics survey software system. First, participants completed electronic informed consent, where, after reading study details, they had the choice of proceeding with the study or quitting. By proceeding they gave their consent for participation. Participants then were asked to fill out a short demographic questionnaire that asked about age, sex, native language, and self-reported English proficiency.

Next, participants were randomly assigned to one of the four experimental conditions in a 2 (Frame: positive, negative) × 2 (Order: certain-option first, uncertain-option first) between-subjects design. They were then presented with a modified financial version of the ADP taken from Tombu and Mandel (2015). Unlike the original ADP, the numeric quantifiers in the two options were qualified with the term *exactly* to increase the likelihood that the two frames would be represented by participants as extensionally equivalent (Mandel, 2014). As well, to reinforce understanding that the numeric quantifiers referred to exact amounts, they were dually described in terms of number and a fraction of the total amount in question. Specifically, in the positive-frame condition, participants were presented with the following description:

Imagine that a financial investment of yours worth $600 has gone sour. If you do nothing you will lose all of it for sure. However, you have two options that are not as bad:

If you choose option A, you will *keep* exactly one-third ($200) of your investment for sure.

If you choose option B, you have exactly a one-third chance of *keeping everything* ($600) and a two-thirds chance of *keeping nothing* ($0).

In the negative-frame condition, the options were alternatively described as follows:

If you choose option A, you will *lose* exactly two-thirds ($400) of your investment for sure.

If you choose option B, you have exactly a one-third chance of *losing nothing* ($0) and a two-thirds chance of *losing everything* ($600).

As in the ADP, the first task was to choose one of the two options. Subsequently, participants were asked how much they preferred their chosen option over the other option on a scale from 1 (no preference) to 7 (strong preference). The choice and preference measures were recoded as follows for the purpose of data analysis. Choices were coded as frame-consistent if and only if they were (a) certain-option choices made in the positive-frame condition or (b) uncertain-option choices made in the negative-frame condition. Next, frame-consistent choices were dummy coded 1, whereas frame-inconsistent choices were dummy coded -1, and these values were multiplied by the strength of preference scores to provide a preference-weighted measure of frame consistency.

After the decision-making task, participants completed the four cognitive-style measures described earlier in the following order: CRT (Frederick, 2005); SNS (Fagerlin et al., 2007); AOT (Baron, 1993; Haran et al., 2013); and NFC (Cacioppo et al., 1984). Using the 3-item version of the CRT, participants were presented with multiple-choice response options. For instance, one problem is, "A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" In earlier studies, participants tend to choose the intuitive answer of 10 cents, even though the correct response is 5 cents. CRT scores were obtained by summing the number of correct responses to the three questions, M_{sum} = 0.69, SD = 1.0, and Cronbach's alpha = 0.71. Participants completed the 8-item SNS (Fagerlin et al., 2007), responding on a 6-point scale (1 = Not at all good, 6 = Extremely good). Examples include, "How often do you find numerical information to be useful?" and "How good are you at figuring out how much a shirt will cost if it is 25% off?" SNS scores were obtained by averaging the responses of the eight items, M = 3.98, SD = 0.94, and Cronbach's alpha = 0.79. AOT was measured using the 7-items used in Haran et al. (2013) and using a 7-point scale (1 = completely disagree, 4 = neutral, and 7 = completely agree). Examples of AOT items include: "People should revise their beliefs in response to new information or evidence" and "Intuition is the best guide in making decisions" (reversed scored). AOT scores were obtained by averaging responses to the seven items, M = 4.88, SD = 0.80, and Cronbach's alpha = 0.67. Finally, we used the 18-item version of the NFC scale (Cacioppo et al., 1984). Participants responded to the item statements on 9-point scales (4 = very strongly agree, 0 = neutral, −4 = very strongly disagree). Examples include, "I find satisfaction in deliberating hard and for long hours" or "It's enough for me that something gets the job done; I don't care how or why it works" (reverse coded). NFC scores were obtained by averaging the responses of the 18-items, M = 5.58, SD = 0.94, and Cronbach's alpha = 0.85.

In general, then, scale reliabilities were above 0.7 (i.e., a conventional cutoff) with the exception of AOT, which was close and not particularly unusual. For instance, Baron et al. (2015) reported a scale reliability of 0.67 for AOT. Haran et al. (2013) did not report scale reliabilities for AOT but Uriel Haran shared the raw data from Experiments 1–3 with us and we verified that the scale reliabilities were 0.70, 0.76, and 0.75 for Experiments 1–3 in Haran et al. (2013), respectively.

We also tested whether any of the cognitive-style measures may have significantly differed across the two framing conditions. Although the differences were non-significant for SNS, AOT, and NFC, CRT scores were significantly higher in the positive-frame condition (M = 0.88, SD = 1.10) than in the negative-frame condition (M = 0.49, SD = 0.86), $t(188) = 2.71$, $p = 0.007$.

Results and Discussion

Out of 190 participants, 118 (62%) made frame-consistent choices (see **Supplementary Data Sheet S1** for data from Experiments 1 and 2). This proportion significantly exceeds chance, as the binomial probability of finding 118 or more frame-consistent choices is 5.21×10^{-4} (in more conventional terms, $p < 0.001$). Recall that frame-consistent choices were coded as 1 and frame-inconsistent choices were coded as −1. The expected value based on chance selection is 0, and the observed mean value is 0.19 (SD = 0.98). By Cohen's (1992) criteria, this corresponds to a small effect size, d = 0.20.

It is of theoretical interest to compare the difference between this effect size estimate and one obtained from a traditional between groups test of the framing effect. Given that the former is more conservative (i.e., it requires choosing the certain option in the positive-frame condition or the uncertain option in the negative-frame condition), we expect that the effect size of the traditional effect will be larger. Indeed, this is the case. Coding selections of the certain and uncertain options as 1 and −1, respectively, reveals a mean value of 0.02 (SD = 1.01) in the negative-frame condition and a mean value of 0.49 (SD = 0.88) in the positive-frame condition, $t(188) = 3.43$, $d = 0.50$, $p = 0.001$. Thus, by applying the stricter (theory-constrained) criterion, the effect size is more than halved.

Recall that Schwitzgebel and Cushman (2015) found that, in line with the reachability bias, frame-consistent choice was more probable when the frame-consistent option was presented first. To test for this effect, we computed a binary measure of whether the frame-consistent choice was presented first or second. The proportion of frame-consistent choices when the frame-consistent option was presented first (0.67) was not significantly greater than the proportion of such choices when the frame-consistent option was presented second (0.58), $p = 0.23$ by Fisher's two-sided exact test. Furthermore, Goodman and Kruskal's tau, which measures the fraction of variability in the categorical variable y (frame-consistent choice) that can be explained by the categorical variable x (whether the frame-consistent option was presented first), was miniscule, $\tau = 0.008$. Therefore, we did not replicate the aforementioned finding by Schwitzgebel and

TABLE 1 | Pearson correlation matrix (Experiment 1).

Variables	1	2	3	4	5	6
1. Frame-consistent choice	1	−0.06	0.16*	−0.04	0.12	0.05
2. Preference strength		1	−0.21*	0.07	0.05	−0.03
3. CRT			1	0.27**	0.29**	0.26**
4. SNS				1	0.21**	0.39**
5. AOT					1	0.32**
6. NFC						1

*$p < 0.05$, **$p < 0.01$, two-tailed.

Cushman (2015)—nor, for that matter, the opposing result of Kühberger and Gradl (2013).

Next, we examined the relations among frame-consistent choice, strength of preference, and the four cognitive-style measures. **Table 1** shows the zero-order (Pearson) correlations. The cognitive-style measures were positively correlated with each other. Frame-consistent choice was significantly correlated with CRT in the predicted direction, but not with any other measure. However, recall that CRT differed across frame. The partial correlation between CRT and frame-consistent choice controlling for frame was not significant, $r(187) = 0.11$, $p = 0.12$.

We followed up the initial correlational analysis by running a binary-logistic regression analysis testing three models. Model 1 includes only the fixed effect (frame), Model 2 further includes the cognitive-style measures, and Model 3 further includes the control variables, sex and age. As **Table 2** shows, the only significant predictor in each of the three models was frame. This result is explained by the fact that the proportion of participants making frame-consistent choices was much larger in the positive-frame condition (0.75) than in the negative-frame condition (0.49). The one-parameter (frame) model was significant, $\chi^2(1, N = 190) = 13.19$, Nagelkerke $R^2 = 0.09$, $p < 0.001$. Model 2 did not significantly improve fit, $\chi^2(4, N = 190) = 7.13$, Nagelkerke $R^2 = 0.14$, $p = 0.13$. Likewise, Model 3 did not improve fit despite an effect of SNS that was almost significant, $\chi^2(2, N = 190) = 1.38$, Nagelkerke $R^2 = 0.15$, $p = 0.50$.

Next, we examined whether preference-weighted frame-consistent choices showed consistent results. Mean preference-weighted choice differed significantly from a test value of zero (as expected by chance) in the frame-consistent direction, $M = 1.09$, 95% CI [0.40, 1.80], one-sample $t(189) = 3.13$, $d = 0.23$, $p = 0.002$. The effect size is comparable to that found in the earlier analysis of unweighted frame-consistent choice. Moreover, this effect size based on the conservative test is, once again, substantially smaller than that obtained by the usual between-groups method. We find a mean value of 0.16 ($SD = 4.95$) in the negative-frame condition and a mean value of 2.28 ($SD = 4.43$) in the positive-frame condition, $t(188) = 3.10$, $d = 0.45$, $p = 0.002$. Finally, consistent with the earlier results of the order-effect analysis on unweighted choice, there was no significant effect of frame-consistent option order on preference-weighted choice, $t(188) = 0.97$, $d = 0.14$, $p = 0.34$.

Finally, we tested a bootstrap multiple linear regression model with frame, NFC, AOT, CRT, SNS, sex, and age as predictors of weighted frame-consistent choice. Model 1 includes only the fixed effect (frame), Model 2 further includes the cognitive-style measures, and Model 3 further includes the control variables, sex and age. The last column of **Table 3** shows that the variance inflation factors (VIFs) are all close to 1, which indicates that the interpretability of the models is not threatened by multicollinearity. **Table 3** shows that, as with the binary-choice measure, only the effect of frame was significant, Model 1 $F(1, 188) = 12.82$, adjusted $R^2 = 0.06$, $p < 0.001$. Model 2 did not significantly improve fit, $F_{change}(4, 184) = 1.48$, $p = 0.21$. Likewise, Model 3 did not improve upon the fit of Model 2, $F_{change}(2, 182) = 0.73$, $p = 0.48$. The results are therefore highly consistent between analyses of binary and preference-weighted choices.

EXPERIMENT 2

The aim of Experiment 2 was to assess the reproducibility of findings from Experiment 1. In most respects, the methods of Experiment 2 were identical to Experiment 1, except that a larger sample of participants was recruited from Amazon Mechanical Turk, and we added one additional cognitive-style measure, Zenhausern's Preference Test (ZPT, Morton, 2002). ZPT was developed in the 1970s as a measure of hemispheric dominance. The 20-item test includes 10 "right hemisphere" (ZPT-R) and 10 "left hemisphere" (ZPT-L) items, with a hemisphericity index scored as the difference of the two subscales (i.e., ZPT-R – ZPT-L). Higher ZPT index values have been interpreted as being indicative of a cognitive disposition toward the use of intuitive System 1 reasoning processes. In particular, McElroy and Seta (2003) found that participants in the highest quartile on the index showed much stronger framing effects than those in the lowest quartile, and they interpreted their findings as supporting the view that frame susceptibility is a cognitive bias owing to reliance on intuitive System 1 reasoning processes. To the best of our knowledge, however, no other study has examined whether this measure predicts frame susceptibility.

Materials and Methods

A sample of 323 Amazon Mechanical Turk (MTurk) workers participated in Experiment 2. The sample was limited to participants who were 18 years of age or older, residing in Canada or the United States, and who completed greater than or equal to 1,000 Human Intelligence Tasks or "HITs" with an approval rate greater than or equal to 95%. Participants were compensated $1.50. Mean age for the sample was 30.4 years ($SD = 6.7$), and 39.2% were female. Seven participants were removed from the analysis because of either (a) unreasonable time for completion (over 24 h, or under 4 min) or (b) reported English proficiency was below 50%. Data from the remaining 316 participants were analyzed.

Except for the change of sample and inclusion of ZPT in the battery of cognitive-style measures, the methods were identical to Experiment 1. Characteristics of the cognitive-style measures were as follows: ZPT-L: $M = 7.36$, $SD = 1.03$, Cronbach's alpha = 0.70; ZPT-R: $M = 5.86$, $SD = 1.46$, Cronbach's alpha = 0.82; CRT: $M = 2.06$, $SD = 1.16$, Cronbach's alpha = 0.80; SNS: $M = 4.66$, $SD = 0.83$, Cronbach's alpha = 0.83; AOT: $M = 5.37$, $SD = 0.93$,

TABLE 2 | Binary logistic regression models predicting frame-consistent choice (Experiment 1).

					95% CI Exp (B)			
Model	**Source**	***B***	***SE***	**Exp(B)**	**LB**	**UB**	**Wald**	***p***
1	Constant	−1.16	0.48	0.31	–	–	5.90	0.015
1	Frame	1.12	0.31	3.05	1.66	5.62	12.79	0.000
2	Constant	−2.01	1.35	0.13	–	–	2.20	0.138
2	Frame	1.11	0.33	3.03	1.60	5.71	11.63	0.001
2	CRT	0.26	0.18	1.30	0.90	1.85	1.99	0.159
2	SNS	−0.31	0.19	0.73	0.50	1.06	2.68	0.102
2	AOT	0.30	0.21	1.35	0.89	2.05	1.97	0.160
2	NFC	0.09	0.19	1.09	0.75	1.59	0.21	0.644
3	Constant	−0.90	1.66	0.41	–	–	0.30	0.587
3	Frame	1.10	0.33	3.01	1.59	5.71	11.44	0.001
3	CRT	0.20	0.19	1.23	0.84	1.78	1.14	0.285
3	SNS	−0.34	0.20	0.71	0.49	1.04	3.05	0.081
3	AOT	0.31	0.22	1.36	0.89	2.08	1.98	0.159
3	NFC	0.15	0.20	1.16	0.78	1.71	0.52	0.472
3	Sex	−0.34	0.40	0.71	0.33	1.54	0.76	0.385
3	Age	−0.03	0.04	0.97	0.90	1.05	0.66	0.417

CI, confidence interval; LB, lower bound; UB, upper bound.

TABLE 3 | Multiple linear regression models predicting preference-weighted frame-consistent choice (Experiment 1).

					95% CI			
Model	**Source**	**β**	***B***	***SE***	**LB**	**UB**	***p***	**VIF**
1	Constant	–	−2.60	1.13	−4.76	−0.24	0.026	–
1	Frame	0.25	2.44	0.67	1.07	3.73	0.001	1.00
2	Constant	–	−4.74	3.05	−11.21	1.24	0.134	–
2	Frame	0.25	2.38	0.67	0.96	3.69	0.001	1.05
2	CRT	0.08	0.36	0.37	−0.42	1.15	0.324	1.20
2	SNS	−0.12	−0.64	0.40	−1.43	0.16	0.105	1.24
2	AOT	0.11	0.68	0.49	−0.37	1.64	0.160	1.19
2	NFC	0.04	0.22	0.45	−0.63	1.17	0.628	1.28
3	Constant	–	−2.35	3.84	−10.85	4.87	0.543	–
3	Frame	0.24	2.36	0.68	0.99	3.60	0.001	1.05
3	CRT	0.05	0.24	0.38	−0.58	1.02	0.528	1.29
3	SNS	−0.13	−0.69	0.40	−1.44	0.10	0.089	1.26
3	AOT	0.11	0.64	0.50	−0.45	1.69	0.194	1.21
3	NFC	0.06	0.29	0.46	−0.63	1.26	0.549	1.38
3	Sex	−0.09	−0.95	0.82	−2.50	0.64	0.257	1.12
3	Age	−0.03	−0.03	0.10	−0.26	0.14	0.731	1.11

All estimates except the standardized regression coefficients are based on 1,000 bias-corrected and accelerated bootstrap samples.

Cronbach's alpha = 0.80; NFC: M = 5.98, SD = 1.62, Cronbach's alpha = 0.95. The ZPT hemisphericity index was computed by subtracting ZPT-L from ZPT-R. All scale reliabilities were good (i.e., >0.70), and invariably greater than in Experiment 1, perhaps reflecting the change of sample from undergraduates to MTurk workers. Finally, none of the measures differed significantly across frame (ps > 0.11).

Results and Discussion

Out of 316 participants, 184 (58.2%) made frame-consistent choices. As in Experiment 1, this proportion significantly exceeds chance: the binomial probability of finding 184 or more frame-consistent choices is 0.002. As noted earlier, the expected value based on chance selection is 0, and the observed mean value is 0.16 (SD = 0.99). This corresponds to a small effect size, d = 0.17, which is very close in magnitude to that found in Experiment 1. By comparison, using the traditional between-groups analysis, we find a mean value of −0.03 (SD = 1.00) in the negative-frame condition and a mean value of 0.30 (SD = 0.96) in the positive-frame condition, $t(314)$ = 3.00, d = 0.34, p = 0.003. Therefore, by applying the stricter, theory-constrained, criterion, the

TABLE 4 | Pearson correlation matrix (Experiment 2).

Variables	1	2	3	4	5	6	7
1. Frame-consistent choice	1	0.04	−0.02	−0.04	0.08	0.10	0.09
2. Preference strength		1	−0.11	−0.01	−0.05	−0.01	0.02
3. CRT			1	0.27**	0.32**	0.19**	−0.22**
4. SNS				1	0.29**	0.32**	−0.10
5. AOT					1	0.34**	−0.20**
6. NFC						1	0.11
7. ZPT							1

*$p < 0.05$, **$p < 0.01$, two-tailed.

TABLE 5 | Binary logistic regression models predicting frame-consistent choice (Experiment 2).

					95% CI Exp (B)			
Model	Source	*B*	*SE*	Exp (B)	LB	UB	Wald	*p*
1	Constant	−0.49	0.36	0.61	–	–	1.88	0.170
1	Frame	0.56	0.23	1.74	1.11	2.74	5.79	0.016
2	Constant	−0.96	0.87	0.38	–	–	1.20	0.273
2	Frame	0.54	0.24	1.71	1.08	2.71	5.19	0.023
2	CRT	−0.02	0.11	0.98	0.79	1.22	0.02	0.876
2	SNS	−0.21	0.15	0.81	0.61	1.08	2.01	0.156
2	AOT	0.19	0.14	1.21	0.92	1.61	1.80	0.180
2	NFC	0.11	0.08	1.12	0.96	1.31	1.92	0.166
2	ZPT	0.13	0.08	1.14	0.97	1.35	2.58	0.108
3	Constant	−0.52	1.07	0.59	–	–	0.24	0.625
3	Frame	0.56	0.24	1.74	1.10	2.78	5.50	0.019
3	CRT	0.00	0.11	1.00	0.80	1.24	0.00	0.998
3	SNS	−0.20	0.15	0.82	0.62	1.10	1.73	0.188
3	AOT	0.23	0.15	1.26	0.94	1.67	2.43	0.119
3	NFC	0.12	0.08	1.12	0.96	1.32	2.02	0.155
3	ZPT	0.15	0.09	1.16	0.98	1.37	2.98	0.084
3	Sex	0.14	0.25	1.15	0.71	1.87	0.33	0.564
3	Age	−0.03	0.02	0.97	0.94	1.01	2.91	0.088

CI, confidence interval; LB, lower bound; UB, upper bound.

effect size was once again substantially reduced (namely, halved).

As in Experiment 1, no order effect was found: the proportion of frame-consistent choices when the frame-consistent option was presented first (0.65) was not significantly greater than the proportion of such choices when the frame-consistent option was presented second (0.55), $p = 0.31$ by Fisher's two-sided exact test, Goodman and Kruskal $\tau = 0.004$.

As **Table 4** shows, and replicating the findings of Experiment 1, frame-consistent choice was not significantly correlated with strength of preference or any of the cognitive-style measures. We followed up the zero-order correlational analysis by running a binary logistic regression analysis. As in Experiment 1, we tested three models: Model 1 included frame only, Model 2 additionally included CRT, SNS, AOT, NFC, and ZPT, and Model 3 additionally included sex and age. As **Table 5** shows, the only significant predictor of frame-consistent choice was frame. As in Experiment 1, the proportion of participants making frame-consistent choices was larger in the positive-frame condition (0.65) than in the negative-frame condition (0.52). The one-parameter (frame) model was significant, $\chi^2(1, N = 316) = 5.85$, Nagelkerke $R^2 = 0.03$, $p = 0.02$. Model 2 did not significantly improve fit, $\chi^2(5, N = 316) = 8.89$, Nagelkerke $R^2 = 0.06$, $p = 0.11$. Likewise, Model 3 did not improve fit over Model 2 in spite of effects of ZPT and age that were almost significant, $\chi^2(2, N = 316) = 3.16$, Nagelkerke $R^2 = 0.07$, $p = 0.21$.

Next, we examined whether preference-weighted choices showed consistent results. Mean preference-weighted choice differed significantly from a test value of zero in the frame-consistent direction, $M = 0.90$, 95% CI [0.32, 1.42], one-sample $t(315) = 3.05$, $d = 0.17$, $p = 0.002$. The effect size for the weighted measure was identical to what we reported earlier for the unweighted measure. Moreover, this effect size based on the conservative test is, once again, substantially smaller than that obtained by the between-groups method. We find a mean value of 0.09 ($SD = 5.09$) in the negative-frame condition and a mean value of 1.89 ($SD = 5.07$) in the positive-frame condition, $t(314) = 3.13$, $d = 0.35$, $p = 0.002$. Finally, showing

TABLE 6 | Multiple linear regression models predicting preference-weighted frame-consistent choice (Experiment 2).

					95% CI			
Model	**Source**	**β**	***B***	***SE***	**LB**	**UB**	***p***	**VIF**
1	Constant	–	−2.07	0.88	−3.81	−0.30	0.021	–
1	Frame	0.19	1.98	0.56	0.86	3.00	0.001	1.00
2	Constant	–	−3.62	2.18	−7.99	1.15	0.103	–
2	Frame	0.18	1.90	0.56	0.72	2.98	0.001	1.02
2	CRT	−0.02	−0.08	0.27	−0.63	0.44	0.725	1.21
2	SNS	−0.07	−0.38	0.38	−1.04	0.29	0.309	1.21
2	AOT	0.07	0.43	0.37	−0.31	1.12	0.246	1.31
2	NFC	0.10	0.31	0.22	−0.09	0.76	0.153	1.27
2	ZPT	0.11	0.39	0.22	−0.04	0.79	0.075	1.13
3	Constant	–	−2.81	2.65	−8.01	2.49	0.298	–
3	Frame	0.18	1.92	0.56	−0.77	2.99	0.001	1.02
3	CRT	−0.01	−0.04	0.27	−0.58	0.47	0.896	1.21
3	SNS	−0.06	−0.34	0.38	−1.01	0.39	0.367	1.22
3	AOT	0.09	0.51	0.37	−0.22	1.19	0.168	1.33
3	NFC	0.10	0.32	0.22	−0.11	0.78	0.150	1.27
3	ZPT	0.12	0.41	0.22	−0.00	0.82	0.062	1.14
3	Sex	0.04	0.46	0.61	−0.77	1.63	0.455	1.05
3	Age	−0.09	−0.07	0.04	−0.15	0.02	0.073	1.04

All estimates except the standardized regression coefficients are based on 1,000 bias-corrected and accelerated bootstrap samples.

strong consistency with the results of Experiment 1, there was no significant effect of frame-consistent option order on preference-weighted choice, $t(314) = 1.05$, $d = 0.12$, $p = 0.30$.

Finally, we tested a bootstrap multiple linear regression model with frame, CRT, SNS, AOT, NFC, ZPT, sex, and age as predictors of weighted frame-consistent choice. The model structure for this analysis was identical to that tested in Experiment 1. As the VIF values shown in the last column of **Table 6** indicate, the interpretability of the models is not threatened by multicollinearity. As **Table 6** shows, Model 1, which includes only frame as a predictor, was significant, $F(1, 314) = 11.77$, adjusted $R^2 = 0.03$, $p = 0.001$. Frame was once again significant in Model 2, and ZPT approached significance, as did the Model 2 improvement of fit, $F_{\text{change}}(5, 309) = 2.04$, adjusted $R^2 = 0.05$, $p = 0.073$. Model 3 did not improve upon the fit of Model 2, $F_{\text{change}}(2, 307) = 1.55$, $p = 0.22$. As in Experiment 1, then, the results were highly consistent between analyses of participants' binary and preference-weighted choices.

GENERAL DISCUSSION

The two online experiments that we conducted yielded highly consistent results despite the fact that one experiment relied on a university undergraduate sample where participants received course credit and the other experiment relied on a MTurk worker sample whose members were paid a nominal rate for their participation. Although it would be advantageous to attempt to replicate the findings in experiments in which participants were not completing the experiment remotely online and in which other variants of the ADP-type task were used, we believe several of the present findings are nevertheless noteworthy.

First, in both experiments we observed levels of frame-consistent decision-making that are unlikely to be due to chance. We find evidence of frame susceptibility even when steps are taken to rule out linguistic interpretations of the options, such as lower bounding of numerical quantifiers, which would invalidate the assumption of extensional equivalence of alternatively framed options. These task design features, which promote clearer interpretability of data from ADP-type tasks (Mandel, 2014), might account for the somewhat lower effect size observed using the conventional between-groups measure. The meta-analytic (framing) effect size for the two experiments reported here is $d = 0.42$, 95% CI [0.22, 0.62]. The meta-analytic effect size from 80 ADP-type studies was 0.57 (Kühberger, 1998). Thus, while lower, the meta-analytic effect size in the present studies is only marginally so.

At first blush, the present findings may appear to contradict those reported by Mandel (2014, Experiment 2). In that experiment, when the term *exactly* was used to prompt a bilateral interpretation of the numerical quantifiers, no significant framing effect was observed. To compare that effect in terms of statistical significance, however, would be misleading because the sample size for that experimental condition was 76, whereas the present experiments matching that condition collectively sampled over 500 participants. Therefore, the present research had much greater statistical power to detect small effects. Drawing on the raw data from the earlier experiment, 44 of the 76 participants (i.e., 57.9%) made frame-consistent choices. The binomial probability of obtaining that number or greater is 0.103. However, the proportion obtained in Mandel (2014, Experiment 2) does not significantly differ from the proportion obtained in Experiments 1 and 2 of the present research. The difference in proportions in the former case (Experiment 1) is 0.042, 95% CI [−0.084,

0.017], and in the latter case (Experiment 2) it is 0.003, 95% CI [−0.115, 0.127]. In other words, quite to the contrary, there is strong consistency in the results, which indicate the existence of a framing effect of small magnitude that may or may not be statistically significant depending on sample size in ADP-type decision-making tasks.

A second noteworthy finding was that no significant effect of option order on choice was found in either experiment. Nevertheless, in each experiment, the proportion of choices that were frame-consistent was greater when the frame-consistent choice was presented first. These non-significant differences are in the same direction as that reported by Schwitzgebel and Cushman (2015), and also in line with the reachability bias (Bar-Hillel et al., 2014). Moreover, if we combine our samples, the effect approaches significance using a one-tailed test: 63.1% choose the frame-consistent choice when that option was presented initially versus 56.1% when that option was presented last, $p = 0.057$ by Fisher's exact one-sided test. Thus, our findings provide faint evidence in support of reachability bias in the context of framing tasks. Although the effect of option order on choice in ADP-type tasks appears to be very weak, it nevertheless should be experimentally controlled (at minimum, through counterbalancing) in future research.

A key finding of this research was that cognitive-style measures had very small predictive effects on frame susceptibility. All were non-significant in each experiment, although there was a small zero-order correlation between CRT and frame susceptibility in Experiment 1 that explained approximately 2.5% of the variance. Taken together, these findings do not support the hypothesis that individual differences in frame susceptibility in decision-making are substantially due to differences in cognitive style—or more specifically, in the degree to which people choose intuitively or deliberately. Moreover, if the true relation between cognitive-style measures and frame susceptibility is weak in the general population, we would expect to see a pattern of results much like we observe in the literature; namely, one in which there appears to be "mixed evidence" in which some studies find significant (but weak) relations and other studies find non-significant relations (that are weak but usually in the expected direction).

Such evidence is "mixed" only in a trivial sense—namely, when researchers pay undue attention to statistical significance across studies that vary in statistical power. The significant effects of cognitive style on frame susceptibility in ADP-type tasks that have been reported in the literature are in most cases small, even when large samples have been used to boost the likelihood of detecting a significant effect (e.g., West et al., 2008). Those results, moreover, are in line with other findings showing that the effect of cognitive ability on judgment and decision-making tasks used to demonstrate cognitive biases and use of heuristic processes is small (e.g., Stanovich and West, 2008; West et al., 2012). The true magnitude of the effect of cognitive style (gauging the System 1/System 2 distinction) on frame susceptibility is therefore likely to lie somewhere between very small and small, using Cohen's (1992) criteria. The precise value is theoretically unimportant because the range is sufficient to indicate that any theory positing that framing effects are largely due to reliance on heuristic "System 1" reasoning processes is wrong. Of course, we do not carelessly generalize this claim to other judgment and decision-making tasks. We acknowledge that there is good evidence that measures of thinking style predict performance on some judgment and decision-making tasks that have been used to demonstrate cognitive biases (e.g., Stanovich and West, 2000). However, this proviso cuts both ways, and we believe researchers should be circumspect in including ADP-type tasks as items in aggregated measures of cognitive bias (e.g., Bruine de Bruin et al., 2007; Toplak et al., 2011).

Critics might counter that we have not made the most of our data by examining the relations between frame-consistent choice and the cognitive-style measures in the combined sample. If we found small but significant effects that would of course reinforce rather than challenge our conclusion. In fact, even with the combined sample of 506 participants, the zero-order correlations between frame-consistent choice and the cognitive-style measures were invariably not statistically significant and the correlations were all close to nil (rs = 0.02, −0.05, 0.08, and 0.08 for CRT, SNS, AOT, and NFC, respectively). Clearly, the results are not due to a lack of statistical power.

Critics might also charge that we have not gone far enough in exploring the possible predictive utility of the measures we investigated. For instance, it is conceivable that CRT would show a stronger relation to frame susceptibility if it were scored in terms of whether the typical intuitive response was selected rather than whether the correct response was selected (Pennycook and Ross, 2016; Stupple et al., 2017). However, this was not the case. If we sum the number of intuitive responses, the correlation remains small in Experiment 1 ($r = -0.14$, $p = 0.056$) and it is virtually nil in Experiment 2 ($r = 0.01$, $p = 0.80$). Nor does an item-response analysis of CRT alter our conclusions. The largest correlation obtained between frame susceptibility and whether or not responses to an item were intuitive was −0.14 (for the lily-pad problem in Experiment 1).

Another possible line of investigation would be to treat the scales as items and to extract factor scores that might prove to be more highly correlated with frame susceptibility. To explore this, we factor analyzed the four measures common to both experiments (CRT, SNS, AOT, and NFC) separately within each experiment. In both cases, using principal components analysis with varimax rotation, a single factor had initial Eigenvalues greater than 1. The factor scores were not significantly correlated with frame susceptibility in either experiment: in Experiment 1, $r = 0.10$ ($p = 0.18$), and in Experiment 2, $r = 0.05$ ($p = 0.41$). Therefore, we find very weak evidence—even using a variety of analytic and data-pooling techniques—to support the hypothesis that individual differences in frame susceptibility are well accounted for by individual differences in thinking style or disposition. To the contrary, the multi-measure, multi-method approach used in this research strongly supports the alternative hypothesis that frame susceptibility in decision-making is not substantially explained by the facets of cognitive style that we examined.

Only one factor explained variation in frame susceptibility in the two experiments and that was the framing manipulation itself. Participants were more likely to make frame-consistent choices in the positive-frame condition than in the negative-frame condition. We strongly suspect that this result is due to a tendency toward risk aversion in the present experiments. This finding is consistent with literature showing that decision-making tasks involving representations of human life (like the ADP) tend to elicit risk-seeking choices, whereas problems with comparable deep structure that instead involve financial outcomes (such as the ADP variant used in the present research) tend to elicit risk-averse choices (e.g., see Jou et al., 1996; Wang, 1996; Fagley and Miller, 1997), perhaps due to the higher aspiration levels set in the morally charged life domain (Schneider, 1992; Rettinger and Hastie, 2003). Hence, the effect of frame on frame susceptibility is likely to be predictable on the basis of content effects on decision-making (Wagenaar et al., 1988; Mandel and Vartanian, 2011). Such content effects, in turn, are likely to be moderated by other decision-task characteristics, such as the payoff structure of choices. In ADP-like problems, a failure to choose would result in maximum sure loss. Clearly (and fortunately), not all decisions are like this. In tasks in which participants must choose between certain and uncertain options but in which inaction implies the status quo, there tends to be greater risk aversion for human-life problems than for monetary problems (Vartanian et al., 2011).

Current theories of framing are not well adapted to explaining such content effects. As noted earlier, most theories of framing make comparable predictions in the ADP—namely, choice of the certain option under positive framing and choice of the uncertain option under negative framing based on inflexible psychophysical assumptions as captured in the stylized value function of prospect theory (Kahneman and Tversky, 1979) or equally inflexible linguistic assumptions as captured in the transformation rules of fuzzy trace theory (Reyna and Brainerd, 1991; Chick et al., 2016). The explicated valence account—or EVA (Tombu and Mandel, 2015), which elucidates how frames (through their explication of outcome valence) affect representations of risk, is more conducive to accommodating content and task effects because the latter, too, appear to influence decision-making through altering risk perceptions. However, EVA currently does not explicitly integrate such factors and would thus require further development.

Much the same could also be said of the editing phase in prospect theory, which is essentially a representational pre-processing stage of decision-making. It is noteworthy that early theoretical attention to framing effects focused on the value function in prospect theory, which predicts risk aversion in the domain of gain and loss aversion in the domain of loss (where the domains are separated by a neutral reference point). Yet, several decades on, it now appears that frames affect the manner in which aspects of problems are mentally represented (Mandel, 2008). The representational effects not only include reference-point selection, as Tversky and Kahneman (1981) had surmised, but also representation of intended communication (e.g., Sher and McKenzie, 2008; van Buiten and Keren, 2009; Teigen, 2011), quantity and probability (Mandel, 2014), and option risk (Tombu and Mandel, 2015). The effects of alternative frames on such representations are probabilistic and naturally give rise to individual differences in representation. For instance, whereas a majority of participants adopted a lower-bound ("at least") interpretation of the certain options in the standard ADP, nearly one-third adopted a bilateral ("exactly") interpretation of the same options (Mandel, 2014, Experiment 3). Surprisingly little research attention has been given to exploring these representational effects. Given how weakly cognitive-style measures predict individual differences in frame susceptibility, research attention to the representational consequences of framing could shed important light on the bases for such individual differences.

AUTHOR CONTRIBUTIONS

DM and IK developed the experiments and analyzed the data. IK executed the experiments. IK contributed to the writing of the manuscript, which was primarily written by DM.

FUNDING

This research was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant 249537-07, Department of National Defence Project 05da, and Canadian Safety and Security Program project 2016-TI-2224.

ACKNOWLEDGMENTS

We thank Uriel Haran for providing us with the raw data reported in Haran et al. (2013) and Jonathan Baron for providing additional information on the Actively Open-minded Thinking scale. We also thank Ulrich Hoffrage and Yasmin Schwegler for their feedback on an earlier draft of this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01461/full#supplementary-material

DATA SHEET S1 | Data for Experiments 1 and 2.

REFERENCES

Aczel, B., Bago, B., Szollosi, A., Foldes, A., and Lukacs, B. (2015). Measuring individual differences in decision biases: methodological considerations. *Front. Psychol.* 6:1770. doi: 10.3389/fpsyg.2015.01770

Almashat, S., Ayotte, B., Edelstein, B., and Margrett, J. (2008). Framing effect debiasing in medical decision making. *Patient Educ. Couns.* 71, 102–107. doi: 10.1016/j.pec.2007.11.004

Arrow, K. J. (1982). Risk perception in psychology and economics. *Econ. Inq.* 20, 1–9. doi: 10.1111/j.1465-7295.1982.tb01138.x

Baldi, P. L., Iannello, P., Riva, S., and Antonietti, A. (2013). Cognitive reflection and socially biased decisions. *Stud. Psychol.* 55, 265–271. doi: 10.21909/sp.2013.04.641

Bar-Hillel, M., Peer, E., and Acquisti, A. (2014). "Heads or tails?"—a reachability bias in binary choice. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 1656–1665. doi: 10.1037/xlm0000005

Baron, J. (1985). "What ends of intelligence components are fundamental?," in *Thinking and Learning Skills*, Vol. 2, eds S. F. Chipman and J. W. Segal (Hillsdale, NJ: Lawrence Erlbaum), 365–390.

Baron, J. (1993). Why teach thinking? An essay. *Appl. Psychol.* 42, 191–214. doi: 10.1111/j.1464-0597.1993.tb00731.x

Baron, J. (2008). *Thinking and Deciding*, 4th Edn, New York, NY: Cambridge University Press.

Baron, J., Scott, S., Fincher, K., and Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *J. Appl. Res. Mem. Cogn.* 4, 265–284. doi: 10.1016/j.jarmac.2014.09.003

Bruine de Bruin, W., Parker, A. M., and Fischhoff, B. (2007). Individual differences in adult decision-making competence. *J. Pers. Soc. Psychol.* 92, 938–956. doi: 10.1037/0022-3514.92.5.938

Cacioppo, J. T., Petty, R. E., and Kao, C. F. (1984). The efficient assessment of need for cognition. *J. Pers. Assess.* 48, 306–307. doi: 10.1207/s15327752jpa4803_13

Campitelli, G., and Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgm. Decis. Mak.* 5, 182–191.

Carnevale, J. J., Inbar, Y., and Lerner, J. S. (2011). Individual differences in need for cognition and decision-making competence among leaders. *Pers. Indiv. Dif.* 51, 274–278. doi: 10.1016/j.paid.2010.07.002

Chick, C. F., Reyna, V. F., and Corbin, J. C. (2016). Framing effects are robust to linguistic disambiguation: a critical test of contemporary theory. *J. Exp. Psychol. Learn. Mem. Cogn.* 42, 238–256. doi: 10.1037/xlm0000158

Cohen, J. (1983). The cost of dichotomization. *Appl. Psychol. Meas.* 7, 249–253. doi: 10.1177/014662168300700301

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159. doi: 10.1037/0033-2909.112.1.115

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.

Covey, J. (2014). The role of dispositional factors in moderating message framing effects. *Health Psychol.* 33, 52–65. doi: 10.1037/a0029305

De Martino, B., Kumaran, D., Seymour, B., and Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science* 313, 684–687. doi: 10.1126/science.1128356

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. (2010). *Thinking Twice: Two Minds in One Brain*. Oxford: Oxford University Press.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., and Smith, D. M. (2007). Measuring numeracy without a math test: development of the Subjective Numeracy Scale (SNS). *Med. Decis. Mak.* 27, 672–680. doi: 10.1177/0272989X07304449

Fagley, N. S., and Miller, P. M. (1997). Framing effects and arenas of choice: your money or your life? *Organ. Behav. Hum. Dec.* 71, 355–373. doi: 10.1006/obhd.1997.2725

Frederick, S. (2005). Cognitive reflection and decision making. *J. Econ. Perspect.* 19, 25–42. doi: 10.1257/089533005775196732

Gamliel, E., Kreiner, H., and Garcia-Retamero, R. (2016). The moderating role of objective and subjective numeracy in attribute framing. *Int. J. Psychol.* 51, 109–116. doi: 10.1002/ijop.12138

Garcia-Retamero, R., and Galesic, M. (2010). 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

Halberg, A.-M., and Teigen, K. H. (2009). Framing of imprecise quantities: when are lower interval bounds preferred to upper bounds? *J. Behav. Decis. Mak.* 22, 490–509. doi: 10.1002/bdm.635

Haran, U., Ritov, I., and Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgm. Decis. Mak.* 8, 188–201.

Humphreys, L. G., and Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. *J. Educ. Psychol.* 66, 464–472. doi: 10.1037/h0036539

Jou, J., Shanteau, J., and Harris, R. J. (1996). An information processing view of framing effects: the role of causal schemas in decision making. *Mem. Cogn.* 24, 1–15. doi: 10.3758/BF03197268

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

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291. doi: 10.2307/1914185

Kee, K. F., and Liang, Y. J. (2015). "Subjective numeracy scale," in *Health Communication Measures*, eds D. K. Kim and J. W. Dearing (New York, NY: Peter Lang), 247–254.

Knight, F. H. (1964). *Risk, Uncertainty, and Profit*. New York, NY: Sentry Press.

Kühberger, A. (1998). The influence of framing on risky decisions: a meta-analysis. *Organ. Behav. Hum. Decis. Process.* 75, 23–55. doi: 10.1006/obhd.1998.2781

Kühberger, A., and Gradl, P. (2013). Choice, rating, and ranking: framing effects with different response modes. *J. Behav. Decis. Mak.* 26, 109–117. doi: 10.1002/bdm.764

LeBoeuf, R. A., and Shafir, E. (2003). Deep thoughts and shallow frames: on the susceptibility to framing effects. *J. Behav. Decis. Mak.* 16, 77–92. doi: 10.1002/bdm.433

Levin, I. P., Gaeth, G. J., Schreiber, J., and Lauriola, M. (2002). A new look at framing effects: distribution of effect sizes, individual differences, and independence of types of effects. *Organ. Behav. Hum. Decis. Process.* 88, 411–429. doi: 10.1006/obhd.2001.2983

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. Mak.* 21, 37–44. doi: 10.1177/0272989X0102100105

Mandel, D. R. (2001). Gain-loss framing and choice: separating outcome formulations from descriptor formulations. *Organ. Behav. Hum. Decis. Process.* 85, 56–76. doi: 10.1006/obhd.2000.2932

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). Do framing effects reveal irrational choice? *J. Exp. Psychol. Gen.* 143, 1185–1199. doi: 10.1037/a0034207

Mandel, D. R., and Vartanian, O. (2011). "Frames, brains, and content domains: neural and behavioral effects of descriptive content on preferential choice," in *Neuroscience of Decision Making*, eds O. Vartanian and D. R. Mandel (New York, NY: Psychology Press), 45–70.

Mata, R., Josef, A. K., Samanez-Larkin, G. R., and Hertwig, R. (2011). Age differences in risky choice: a meta-analysis. *Ann. N. Y. Acad. Sci.* 1235, 18–29. doi: 10.1111/j.1749-6632.2011.06200.x

Maxwell, S. E., and Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychol. Bull.* 113, 181–190. doi: 10.1037/0033-2909.113.1.181

Mayhorn, C. B., Fisk, A. D., and Whittle, J. D. (2002). Decisions, decisions: analysis of age, cohort, and time of testing on framing of risky decision options. *Hum. Factors* 44, 515–521. doi: 10.1518/0018720024496935

McElroy, T., and Seta, J. J. (2003). Framing effects: an analytic–holistic perspective. *J. Exp. Soc. Psychol.* 39, 610–617. doi: 10.1016/S0022-1031(03)00036-2

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., et al. (2015). The psychology of intelligence analysis: drivers of prediction accuracy in world politics. *J. Exp. Psychol. Appl.* 21, 1–14. doi: 10.1037/xap0000040

Miller, P. M., and Fagley, N. S. (1991). The effects of framing, problem variations, and providing rationale on choice. *Pers. Soc. Psychol. Bull.* 17, 517–522. doi: 10.1177/0146167291175006

Morton, B. E. (2002). Outcomes of hemisphericity questionnaires correlate with unilateral dichotic deafness. *Brain Cogn.* 49, 63–72. doi: 10.1006/brcg.2001.1485

Noori, M. (2016). Cognitive reflection as a predictor of susceptibility to behavioral anomalies. *Judgm. Decis. Mak.* 11, 114–120.

Oechssler, J., Roider, A., and Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *J. Econ. Behav. Organ.* 72, 147–152. doi: 10.1016/j.jebo.2009.04.018

Pennycook, G., and Ross, R. M. (2016). Commentary: cognitive reflection vs. calculation in decision making. *Front. Psychol.* 7:9. doi: 10.3389/fpsyg.2016.00009

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., and Bjalkebring, P. (2015). Multiple numeric competencies: when a number is not just a number. *J. Pers. Soc. Psychol.* 108, 802–822. doi: 10.1037/pspp0000019

Peters, E., Hart, P. S., and Fraenkel, L. (2011). Informing patients: the influence of numeracy, framing, and format of side effect information on risk perceptions. *Med. Decis. Mak.* 31, 432–436. doi: 10.1177/0272989X10391672

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., 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

Piñon, A., and Gambara, H. (2005). A meta-analytic review of framing effect: risky, attribute and goal framing. *Psicothema* 17, 325–331.

Rettinger, D. A., and Hastie, R. (2003). "Comprehension and decision making," in *Emerging Perspectives on Judgment and Decision Research: Cambridge Series on Judgment and Decision Making*, eds S. L. Schneider and J. Shanteau (New York, NY: Cambridge University Press), 165–200. doi: 10.1017/CBO9780511609978.008

Reyna, V. F., and Brainerd, C. J. (1991). Fuzzy-trace theory and framing effects in choice: gist extraction, truncation, and conversion. *J. Behav. Decis. Mak.* 4, 249–262. doi: 10.1002/bdm.3960040403

Rönnlund, M., Karlsson, E., Laggnäs, E., Larsson, L., and Lindström, T. (2005). Risky decision making across three arenas of choice: are younger and older adults differently susceptible to framing effects? *J. Gen. Psychol.* 132, 81–93. doi: 10.3200/GENP.132.1.81-93

Schneider, S. L. (1992). Framing and conflict: aspiration level contingency, the status quo, and current theories of risky choice. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 1040–1057. doi: 10.1037/0278-7393.18.5.1040

Schwitzgebel, E., and Cushman, F. (2015). Professional philosophers' susceptibility to order effects and framing effects in evaluating moral dilemmas. *Cognition* 141, 127–137. doi: 10.1016/j.cognition.2015.04.015

Sher, S., and McKenzie, C. R. M. (2008). "Framing effects and rationality," in *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, Vol. 79–96, eds N. Chater and M. Oaksford (Oxford: Oxford University Press).

Sieck, W., and Yates, J. F. (1997). Exposition effects on decision making: choice and confidence in choice. *Organ. Behav. Hum. Decis. Process.* 70, 207–219. doi: 10.1006/obhd.1997.2706

Simon, A. F., Fagley, N. S., and Halleran, J. G. (2004). Decision framing: moderating effects of individual differences and cognitive processing. *J. Behav. Decis. Mak.* 17, 77–93. doi: 10.1002/bdm.463

Sinayev, A., and Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Front. Psychol.* 6:532. doi: 10.3389/fpsyg.2015.00532

Smith, S. M., and Levin, I. P. (1996). Need for cognition and choice framing effects. *J. Behav. Decis. Mak.* 9, 283–290. doi: 10.1111/j.1467-8519.2012.01973.x

Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behav. Brain Sci.* 23, 645–665. doi: 10.1017/S0140525X00003435

Stanovich, K. E., and West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *J. Pers. Soc. Psychol.* 94, 672–695. doi: 10.1037/0022-3514.94.4.672

Strough, J., Karns, T. E., and Schlosnagle, L. (2011). Decision-making heuristics and biases across the life span. *Ann. N. Y. Acad. Sci.* 1235, 57–74. doi: 10.1111/j.1749-6632.2011.06208.x

Stupple, E. J. N., Pitchford, M., Ball, L. J., Hunt, T. E., and Steel, R. (2017). Slower is not always better: response-time evidence clarifies the limited role of miserly information processing in the cognitive reflection test. *PLoS One* 12:e0186404. doi: 10.1371/journal.pone.0186404

Szaszi, B., Szollosi, A., Palfi, B., and Aczel, B. (2017). The cognitive reflection test revisited: exploring the ways individuals solve the test. *Think. Reason.* 23, 207–234. doi: 10.1080/13546783.2017.1292954

Takemura, K. (1994). Influence of elaboration on the framing of decision. *J. Psychol.* 128, 33–39. doi: 10.1080/00223980.1994.9712709

Teigen, K. H. (2011). "When frames meet realities: on the perceived correctness of inaccurate estimates," in *Perspectives on Framing*, ed. G. Keren (London: Psychology Press), 197–217.

Teigen, K. H., and Nikolaisen, M. I. (2009). Incorrect estimates and false reports: how framing modifies truth. *Think. Reason.* 15, 268–293. doi: 10.1080/13546780903020999

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?*. Princeton, NJ: Princeton University Press.

Tombu, M., and Mandel, D. R. (2015). When does framing influence preferences, risk perceptions, and risk attitudes? The explicated valence account. *J. Behav. Decis. Mak.* 28, 464–476. doi: 10.1002/bdm.1863

Toplak, M. E., and Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: searching for a generalizable critical thinking skill. *J. Educ. Psychol.* 94, 197–209. doi: 10.1037/0022-0663.94.1.197

Toplak, M. E., West, R. F., and Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Mem. Cogn.* 39, 1275–1289. doi: 10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., and Stanovich, K. E. (2014). Assessing miserly information processing: an expansion of the cognitive reflection test. *Think. Reason.* 20, 147–168. doi: 10.1080/13546783.2013.844729

Toplak, M. E., West, R. F., and Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample. *J. Behav. Decis. Mak.* 30, 541–554. doi: 10.1002/bdm.1973

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science* 211, 453–458. doi: 10.1126/science.7455683

Tversky, A., and Kahneman, D. (1986). Rational choice and the framing of decisions. *J. Bus.* 59, S251–S278. doi: 10.1086/296365

van Buiten, M., and Keren, G. (2009). Speaker-listener incompatibility: joint and separate processing in risky choice framing. *Organ. Behav. Hum. Decis. Process.* 108, 106–115. doi: 10.1016/j.obhdp.2008.03.002

Vartanian, O., Mandel, D. R., and Duncan, M. (2011). Money or life: behavioral and neural context effects on choice under uncertainty. *J. Neurosci. Psychol. Econ.* 4, 25–36. doi: 10.1037/a0021241

Wagenaar, W. A., Keren, G., and Lichtenstein, S. (1988). Islanders and hostages: deep and surface structures of decision problems. *Acta Psychol.* 67, 175–189. doi: 10.1016/0001-6918(88)90012-1

Wallin, A., Paradis, C., and Katsikopoulos, K. V. (2016). Evaluative polarity words in risky choice framing. *J. Pragmat.* 106, 20–38. doi: 10.1016/j.pragma.2016.09.005

Wang, X. T. (1996). Framing effects: dynamics and task domains. *Organ. Behav. Hum. Decis. Process.* 68, 145–157. doi: 10.1006/obhd.1996.0095

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., and Peters, E. (2013). Development and testing of an abbreviated numeracy scale: a rasch analysis approach. *J. Behav. Decis. Mak.* 26, 198–212. doi: 10.1002/bdm.1751

West, R. F., Meserve, R. J., and Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *J. Pers. Soc. Psychol.* 103, 506–519. doi: 10.1037/a0028857

West, R. F., Toplak, M. E., and Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: associations with cognitive ability and thinking dispositions. *J. Educ. Psychol.* 100, 930–941. doi: 10.1037/a0012842

Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., and Fagerlin, A. (2007). Validation of the subjective numeracy scale: effects of low numeracy on comprehension of risk communications and utility elicitations. *Med. Decis. Mak.* 27, 663–671. doi: 10.1177/0272989X07303824

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: 09 October 2018
doi: 10.3389/fpsyg.2018.01906

Too Worried to Judge: On the Role of Perceived Severity in Medical Decision-Making

Àngels Colomé[1,2], Javier Rodríguez-Ferreiro[1,2] and Elisabet Tubau[1,2]*

[1] Section of Cognitive Processes, Department of Cognition, Development and Educational Psychology, Faculty of Psychology, University of Barcelona, Barcelona, Spain, [2] Institute of Neurosciences, University of Barcelona, Barcelona, Spain

Ideally, decisions regarding one's health should be made after assessing the objective probabilities of relevant outcomes. Nevertheless, previous beliefs and emotional reactions also have a role in decision-making. Furthermore, the comprehension of probabilities is commonly affected by the presentation format, and by numeracy. This study aimed to assess the extent to which the influence of these factors might vary between different medical conditions. A sample of university students were presented with two health scenarios containing statistical information on the prevalence of breast cancer and hypertension either through icon arrays (N = 71) or natural frequencies (N = 72). They also received information regarding a preventive measure (mammogram/low-sodium diet) and the likelihood of a positive mammogram or a rich-sodium diet either when suffering or not suffering from the disease. Before seeing the data, participants rated the severity of the disease and the inconvenience of the preventive measure. After reading the health scenario, participants had to rate its difficulty, and how worrisome it was. They had also to rate the prior probability of suffering from this medical condition, and the posterior probability of it, provided a positive mammogram or a rich-sodium diet. Finally, they rated the extent to which they would recommend the preventive measures. All the rates used the same 1 (little)-8 (a great deal) scale. Participants' numeracy was also assessed. The scenarios differed significantly in perceived severity and worry, with the cancer scenario obtaining higher scores. Importantly, regression analyses showed that the recommendations in the two health scenarios depended on different variables. A model taking into consideration severity and worry rates best explained decisions in the cancer scenario; in contrast, in the hypertension scenario the model that best explained the recommendations comprised both the posterior probability estimate and the severity rate. Neither numeracy nor presentation format affected recommendation but both affected difficulty, worrying and probability rates. We conclude that previous perceptions of the severity of a health condition modulate the use of probabilistic information for decision-making. The roles of presentation format and numeracy in enabling patients to understand statistical information are also discussed.

Keywords: medical decision, severity perception, probability judgment, affect, numerical format, numeracy

OPEN ACCESS

Edited by:
Gorka Navarrete,
Adolfo Ibáñez University, Chile

Reviewed by:
Dafina Petrova,
Andalusian School of Public Health, Spain
Alistair Andres Thorpe,
University of Essex, United Kingdom

****Correspondence:***
Àngels Colomé
angels.colome@ub.edu

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

Received: *23 April 2018*
Accepted: *18 September 2018*
Published: *09 October 2018*

Citation:
Colomé À, Rodríguez-Ferreiro J and Tubau E (2018) Too Worried to Judge: On the Role of Perceived Severity in Medical Decision-Making. Front. Psychol. 9:1906. doi: 10.3389/fpsyg.2018.01906

INTRODUCTION

Passing from a doctor-centered to a patient-centered model of health has led in the last decades to an increase in the interest devoted to informed consent and how to ensure that decisions are indeed knowledgeable. Informed consent should be provided after the patient has understood the purpose, benefits and potential risks of the alternatives proposed. Risks are often conceptualized as a combined function of the probability of a loss and its consequences (Lipkus, 2007). Hence, in health contexts, risk assessment will depend on its probability but also on how severe this risk is considered to be (Haase et al., 2013). Although the probability and the subjective value of the outcome are usually assumed to be independent constructs, Harris et al. (2009) showed that this was not always the case. They found a main effect of probability but, interestingly, estimation at each probability level was higher when the consequences of participants' decisions were more severe. Harris et al. attributed their effect to the fact that, in case of severe consequences, the costs associated with underestimating probability are high; individuals, therefore, inflate their estimations of the probability of occurrence as a preventive measure. However, this would happen only when participants can make a decision based on these probabilities.

Importantly, Harris et al. (2009) suggested that the effects of outcome severity would be larger under conditions of emotional involvement. If this was the case, understanding how patients make their medical decisions might require an assessment of their comprehension of the objective information conveyed, but also a consideration of how they interpret it on the basis of their background (e.g., their previous perceptions of the disease and remediation proposed or their attitudes toward them) as well as their affect with regard to it.

Affect has been defined as the "specific quality of "goodness" or "badness" (a) experienced as a feeling state and (b) demarcating a positive or negative quality of a stimulus" (p. 312, Slovic et al., 2004). Finucane et al. (2000) considered that people may base their judgments of an item not only on what they thought about it, but also on how they felt about it, and coined the term "affect heuristic" to name this phenomenon. Loewenstein et al. (2001) talked for the first time of the importance of anticipatory emotions, i.e., immediate visceral reactions to risk and uncertainty such as worry or anxiety, and proposed the "risk-as-feelings hypothesis". According to these authors, apparently erratic decisions might be due to the fact that people's emotional reactions to risk respond to factors other than the cognitive evaluation of risks, and are largely insensitive to differences in probability. Finally, other studies such as Pachur et al. (2014) have shown that individuals behave differently in affect-rich (e.g., concerning the side effects of a drug) and affect-poor (monetary) contexts which are otherwise equivalent. Pachur et al. (2014) concluded that affect acted as a "spotlight", focusing people's attention on outcomes and leading them to neglect statistical information.

Our aim in this study was to investigate whether previous beliefs and affects related to the severity of a given medical condition and a possible preventive measure might influence the extent to which participants would recommend a loved one or friend to use this measure. Furthermore, we wondered whether these factors might affect the way they process the probability information conveyed. In contrast with previous research, perceived severity and the inconvenience caused by the preventive measure were assessed before exposure to the information in order to ensure that our participants' responses were not influenced by the data provided.

In addition to previous beliefs, perceived severity and associated emotional reactions might also depend on the format in which numerical information is presented. In a previous study, we found that representing frequencies in the form of icon arrays makes them easier to understand than presenting them as Arabic digits, especially when having to infer posterior probabilities (Tubau et al., 2018).[1] However, Petrova et al. (2015) concluded that visual aids only helped people for whom the medical information provided was not too affectively imbued; in contrast, people seeing the disease as extremely unpleasant or severe did not pay attention to the statistical information provided, and made their decision based on their previous beliefs of the effectiveness of screening or their fear of the disease. Also in the context of medical scenarios but with a different approach, Timmermans et al. (2008) found that human icons had more affective impact than frequencies or percentages, and risks presented as icons were judged as more likely. Nevertheless, format affected the decision in just one out of their four scenarios and some uncontrolled features of the scenarios make it difficult to extract general conclusions on the relationship between affective response and the intention to recommend preventive measures. All in all though, previous evidences suggested that presentation format was a variable to take into account.

Finally, it is worth noting that the effects of previous beliefs and affect might be also modulated by individual level of numeracy. Numeracy is defined as "the ability to process basic probability and numerical concepts" (Peters et al., 2006, p. 407). People with low numeracy are not only less accurate in estimating probabilities than their high in numeracy peers, but also more prone to frame, text complexity and numerical format effects (e.g., Peters et al., 2006, 2011; Johnson and Tubau, 2013). Furthermore, previous studies have found differences between people with low and high numeracy in both risk perception and commitment to take certain decisions, with people with low numeracy being less able to integrate probabilities and outcome information, particularly in affect-rich contexts (e.g., Pachur and Galesic, 2012). In contrast, emotions of people with high numeracy vary more in proportion to the probability of the loss than their peers (Petrova et al., 2014). Given these previous data, we decided to assess the numeracy of our participants by asking them to answer a selected sample of the items in the numeracy scale by Lipkus et al. (2001; see section Materials and Procedure).

Participants in our experiment were presented with two medical scenarios, one concerning breast cancer and the other regarding hypertension. These two medical conditions were

[1]Most of previous studies using icons presented them together with numerical frequencies and, perhaps because of this, mixed findings on format effects have been reported (see for example the meta-analysis by McDowell and Jacobs, 2017). Hence, to avoid potential confounds, our study used icons without the redundant numerical information.

selected because they were expected to differ in their perceived severity. Our sample consisted of university students, mainly women, in their early twenties: we hypothesized that, even though hypertension is more prevalent than breast cancer and has well-known possible negative consequences (e.g., a higher likelihood of suffering an ictus or heart attack), it would not be considered as lethal *per se*, especially by the sample in question. In order to verify our hypothesis and assess our participants' previous beliefs, we explored how severe they considered the two medical conditions to be, before presenting them with any prevalence data. We also asked them about their beliefs regarding the two preventive measures they would have to recommend.

Subsequently, the two medical scenarios were proposed. Both included information on the prevalence of the disease, as well as data on a preventive measure. Health care campaigns often stress the positive effects of preventive measures and tend to omit the bothersome or even negative consequences of their use such as overdiagnosis. As a result, people may be well disposed to use them, even without considering the information provided (Petrova et al., 2016). So, in order to avoid an indiscriminate "yes" response to the recommendation, both health scenarios included also a drawback of it.

After the presentation of the medical scenario, we asked our participants again about the affect (worry) that the current information had aroused in them. We also wanted to determine how difficult they found it to understand the information provided. Finally, we asked them to rate the prior and posterior probabilities (see method) and to decide whether they would recommend this remediation measure. Our predictions were as follows. We expected that, prior to testing, participants would view breast cancer as more severe than hypertension. As for the perceived inconvenience of preventive measures, we did not have any preconceptions: we merely wanted to measure participants' previous beliefs and feelings. Regarding the subsequent items, we expected that information on the more severe disease would also be considered as more worrying. We also predicted that, although participants would take into account the likelihood of the events, the weight of numerical information on the decision process might depend on the scenario: we expected that higher levels of worry and severity would make participants more likely to recommend preventive measures above and beyond the perceived probabilities.

The statistical data for each scenario were presented either verbally, with quantities reported as natural frequencies in Arabic numerals, or through arrays of 100 icons (see **Figure 1**). Regarding the format, we aimed to test two alternative hypotheses. On the one hand, according to Timmermans et al. (2008), higher vividness of the risks displayed as icons should cause more affective response in participants than digits; this should increase the perceived probability and the commitment to recommend the preventive measure, especially in more severe medical situations. We considered that this effect might be maximized by the use of anthropomorphic figures, so we used restroom-like icons. On the other hand, based on the above mentioned benefit of icons for risk comprehension, we expected more sensitivity to the probability information for the ratings in this format.

Regarding the effect of numeracy, we hypothesized that people with low numeracy would consider information to be harder to understand than their high in numeracy peers, but they might also see it as more worrying and more likely to occur. This, in turn, might translate into a higher intention to recommend the preventive measure, especially in the more severe medical condition. In contrast, high-skilled participants might adjust their recommendation more to the probability ratings.

In sum, two scenarios differing in severity were used to investigate whether previous beliefs and affect related to a given medical condition and a possible preventive measure might influence the extent to which participants would recommend to use this measure. Given previous evidences of the relevance of these two variables in probability processing and decision making, format was manipulated and numeracy of participants was measured.

MATERIALS AND METHODS

Participants

One hundred and forty-three Psychology students [115 women and 28 men, mean age = 23.37 (SD = 5.98)] from the University of Barcelona took part in this experiment as part of their course. Sensitivity analysis conducted with GPower (Faul et al., 2007) shows that for our main variable of interest, i.e., severity of the medical scenario, this sample size implies a minimal detectable effect of f = 0.15, which is considered to be small according to Cohen (1992).

Probabilistic reasoning and the Bayes rule were introduced only after this session. Participants were free to join in the experiment, and provided written consent for the use of their data for research purposes. They were debriefed in a subsequent session.

Materials and Procedure

Participants were presented with two health scenarios concerning breast cancer and hypertension. Each scenario ended with the possibility of using a preventive measure (mammogram/low-sodium diet). Information was presented through icon arrays (see **Figure 1**) to 71 and in the form of natural frequencies (e.g., "3 of the 4 women with breast cancer and 12 of the 96 women without breast cancer receive a positive mammogram") to the rest (N = 72). Participants were randomly assigned to each format condition. They were tested collectively, although each one had their own computer and solved the task individually. There were no time limitations for answering, although it took all participants between 15 and 20 min to complete the whole task.

The procedure was as follows[2]. An initial screen informed participants they would receive some data concerning the prevalence of breast cancer, and asked them to rate how severe they considered this disease and how inconvenient they thought mammograms were. Participants were required to respond using a 1-to-8 scale, on which a score of 1 meant hardly

[2]For the sake of clarity, here we describe one particular session; however, the order of the two scenarios was counterbalanced across participants.

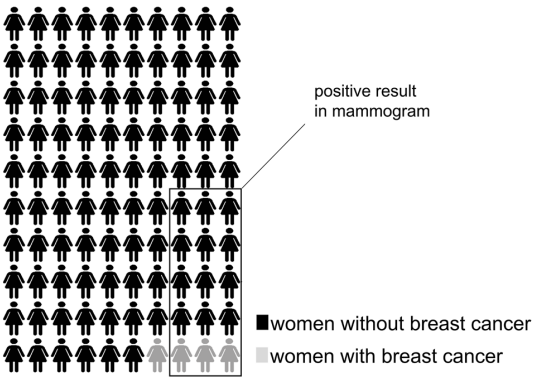


FIGURE 1 | Iconic representation of the breast cancer problem (original version in Spanish and in color).

severe/inconvenient at all and 8 highly severe/inconvenient. A second screen reported the prevalence of breast cancer in women over 50, and the reliability of mammograms for its detection (prevalence: 4 of 100; hit rate: 3 of 4; false alarm rate: 12 of 96). It also noted that a positive mammogram result (true or false) might require a needle biopsy to confirm it. In order to avoid effects due to differences in individual memory, these data remained on the screen until participants had answered all the items on this medical scenario. Questions on this screen asked participants how difficult they had found it to understand the probability of suffering from breast cancer (1 = very easy; 8 = very difficult) and how worrisome (1 = not worrisome at all; 8 = very worrisome) they found the information provided. A third screen asked participants to think about a friend or relative in this age group and to rate the probability that she might suffer from breast cancer (prior probability; 1 = very unlikely; 8 = very likely), or that she might suffer from breast cancer if she had received a positive result in the mammogram (posterior probability). Finally, participants were requested to rate the extent to which they would recommend the mammogram to their friend or relative (1 = definitely not; 8 = definitely).

In the hypertension scenario, participants first had to rate the severity of the condition and the inconvenience of following a low-sodium diet. After this, they were provided with the base rate of women over 40 suffering from this medical condition as well as the rate of women following a sodium-rich diet with or without hypertension (prevalence: 20 of 100; hit rate: 12 of 20; false alarm rate: 24 of 80). They were also reminded that doctors often recommend a low-sodium diet, even though many people consider it to be unpleasant. Much as in the previous scenario, participants were required to rate the difficulty of understanding the information provided and how worrying it was (screen 2); the probability that a friend or relative of this age might suffer from hypertension, and the same probability if she followed a sodium-rich diet (screen 3). Finally, participants were told that their friend or relative was considering following a low-sodium diet and they were asked to decide whether they would recommend it.

We also assessed participants' numeracy using the four items (see **Appendix**) rated by Peters et al. (2006) as the most difficult ones on the numeracy scale of Lipkus et al. (2001). Three of these items were the ones previously used by Schwartz et al. (1997). Participants answered these questions at the end of the session.

RESULTS

We had hypothesized that the two scenarios would differ in the previous beliefs and affect aroused by the medical condition and remediation presented and that this might have consequences in the likelihood of recommendations. Hence, our first analysis was devoted to confirm the existence of differences between the two medical scenarios. Given that we predicted that format and numeracy might also have an effect on the comprehension of the data and the affect aroused by them, we also entered these two variables into the analysis. Nevertheless, for the sake of comprehension, we will report the data concerning them separately.

We conducted an ANOVA for each dependent variable (responses to each question) with the medical scenario (breast cancer and hypertension) as a within-participant variable, and format (icons and natural frequencies) and numeracy (low and high) as between-participant factors. As for numeracy, participants were classified into two groups according to their performance on the numeracy questionnaire: they were considered as showing low numeracy (LN) if they had correctly answered two items or fewer (N = 69) and as having high numeracy (HN) if they had correctly answered three or four (N = 74).

Effects of Scenario

For the sake of readability, F values, significance and descriptive statistics are presented in **Table 1**. As expected, the effect of scenario was significant for *perceived severity*, *worry*, *prior* and *posterior probability* and *recommendation rates*. For worry and probability ratings, scenario also interacted with format and numeracy (see below). More specifically, participants judged breast cancer to be significantly more severe than hypertension. The breast cancer scenario also raised more worry than the hypertension scenario, and the mammogram was recommended significantly more frequently than the low-sodium diet (see **Table 1**). Even if it was not theoretically relevant, finding no effects of scenario in difficulty helped us discarding that differences between scenarios were due to problems in comprehending one of them. As for probabilities, the fact that the ratings of the prior and posterior probabilities differed across scenarios, with higher ratings for hypertension, indicates that participants' answers were sensitive to the disparity in the numerical information provided in each of them.

Effects of Format

As above, here we report only significant effects. See **Table 2** for a detailed list of the descriptive statistics as well as F and p values.

Format affected *difficulty*, *worry* (in interaction with scenario and numeracy; see below), and *posterior probability* rates (in interaction with scenario). That is, the data presented through icons were always judged to be *easier* to understand than data

TABLE 1 | Effects of the medical condition (severity of the scenario).

	Breast Cancer	Hypertension			
	Mean *(SD)*	**Mean *(SD)***	***F* values**	***p***	η_p^2
Perceived severity	7.20 (0.96)	5.89 (1.20)	$F(1,139) = 160.40$	<0.001	0.53
Inconvenience measures	4.17 (2.13)	4.09 (2.13)	$F(1,139) < 1$		
Difficulty	3.56 (1.87)	3.27 (1.64)	$F(1,139) = 3.17$	0.07	0.022
Worry	5.14 (1.68)	4.83 (1.60)	$F(1,139) = 5.17$	0.024	0.036
Prior probability	2.64 (1.48)	3.50 (1.51)	$F(1,139) = 43.52$	<0.001	0.23
Posterior probability	4.14 (1.89)	5.03 (1.71)	$F(1,139) = 27.88$	<0.001	0.16
Recommendation	7.34 (1.06)	5.60 (1.83)	$F(1,139) = 117.95$	<0.001	0.45

Descriptive statistics [Mean and Standard Deviation (SD)] and comparison of responses to the breast cancer and hypertension medical scenarios.

TABLE 2 | Effects of format.

	Breast Cancer		Hypertension				
	Icons	**Frequencies**	**Icons**	**Frequencies**			
	Mean *(SD)*	**Mean *(SD)***	**Mean *(SD)***	**Mean *(SD)***	***F* values**	***p***	η_p^2
Perceived severity	7.21 (.89)	7.19 (0.94)	5.74 (1.10)	6.04 (1.28)			
Inconvenience	4.15 (2.24)	4.18 (2.04)	4.12 (2.22)	4.05 (2.05)			
Difficulty	3.04 (1.76)	4.07 (1.84)	2.89 (1.59)	3.65 (1.62)	Format: $F(1,139) = 12.75$	< 0.001	0.084
Worry	5.24 (1.66)	5.04 (1.71)	4.99 (1.61)	4.67 (1.58)	Scenario x Numeracy x Format: $F(1,139) = 3.88$	0.051	0.027
Prior probability	2.49 (1.39)	2.78 (1.56)	3.52 (1.47)	3.47 (1.56)			
Posterior probability	3.89 (1.79)	4.39 (1.96)	5.20 (1.75)	4.86 (1.67)	Scenario x Format: $F(1,139) = 6.29$	0.013	0.043
Recommendation	7.35 (1.10)	7.34 (1.02)	5.83 (1.82)	5.38 (1.83)			

Descriptive statistics [Mean and Standard Deviation (SD)] of the responses to each question and summary of results of the ANOVA investigating the differences between icons and natural frequencies.

presented through written frequencies. Furthermore, participants who received the information in iconic format were more sensitive to variation in probabilities than those who saw it as frequencies: only the *posterior probability ratings based on icons* were correctly identified as differing across the scenarios [$t(70) = 5.73$, $p < 0.001$ and $t(71) = 1.86$; $p = 0.06$ for scenarios presenting icons and natural frequencies, respectively; see **Table 2**].

Effects of Numeracy

Numeracy showed significant effects for *difficulty*, *worry* (in interaction with scenario and format) and *prior probability* rates (see **Table 3**). As expected, people with low numeracy rated the information provided in both scenarios as *more difficult* to comprehend than those scoring high in numeracy. They also rated breast cancer as *more worrying* than people with high numeracy when the data were presented through icons (means of worry rates were 4.8 and 5.8 for HN and LN, respectively; $t(69) = 2.63$, $p = 0.01$), but not in the case of natural frequencies (mean of worry rate in either group was 5; $t < 1$). Last, participants with low numeracy judged *breast cancer to be more likely* than their high in numeracy peers (see **Table 3**).

Factors Influencing Recommendation

Our second analysis aimed at determining which variables might have affected decisions in a particular medical scenario. We first conducted a correlational analysis to check which variables (numeracy, format, scores on the items concerning disease severity, inconvenience caused by the preventive measure, difficulty to comprehend and worrying, as well as the estimated prior and posterior probabilities) significantly correlated with the likelihood to recommend the preventive measure. Subsequently we conducted a forced entry multiple regression for each scenario introducing the significant variables in the correlation analyses as potential predictors and using the scores in the recommendation item as dependent variable.

Scenario 1. Breast Cancer

Commitment to recommend correlated significantly with disease severity and worry (see **Table 4**). A model including these two variables accounted for 12% of the variance, $R^2 = 0.12$, adjusted $R^2 = 0.11$; $F(2,140) = 10.17$, $p < 0.001$. Disease severity and degree of worrying were both significant predictors of participants' recommendation ($\beta = 0.29$, $p < 0.001$ and $\beta = 0.17$, $p = 0.02$, respectively) with disease severity receiving more weight.

Scenario 2. Hypertension

Commitment to recommend correlated significantly with severity, worry, prior probability and posterior probability (see **Table 5**). A model comprising these variables reached significance, $F(4,138) = 6.90$, $p < 0.001$, and explained 16% of the variance in the recommendation of participants: $R^2 = 0.16$, adjusted $R^2 = 0.14$. When looking at each particular predictor, only severity and posterior probability reached significance

TABLE 3 | Effects of numeracy.

	Breast Cancer		Hypertension				
	Low numerates	High numerates	Low numerates	High numerates			
	Mean *(SD)*	Mean *(SD)*	Mean *(SD)*	Mean *(SD)*	*F* values	*p*	η_p^2
Perceived severity	7.23 (0.91)	7.18 (0.92)	6.01 (1.26)	5.78 (1.13)			
Inconvenience	4.33 (2.25)	4.01 (2.02)	3.60 (1.99)	4.54 (2.17)	Scenario × Numeracy: $F(1,139) = 7.07$	0.009	0.048
Difficulty	4.03 (1.86)	3.12 (1.79)	3.80 (1.70)	2.78 (1.43)	Numeracy: $F(1,139) = 14.65$	<0.001	0.095
Worry	5.39 (1.69)	4.91 (1.65)	4.74 (1.74)	4.91 (1.46)	Scenario × Numeracy: $F(1,139) = 4.43$	0.037	0.031
					Scenario × Numeracy × Format: $F(1,139) = 3.88$	0.051	0.027
Prior probability	3.13 (1.74)	2.18 (1.01)	3.64 (1.57)	3.36 (1.45)	Numeracy: $F(1,139) = 8.35$	0.004	0.057
					Scenario × Numeracy: $F(1,139) = 6.41$	0.012	0.044
Posterior probability	4.32 (1.96)	3.97 (1.81)	5.25 (1.73)	4.82 (1.68)			
Recommendation	7.27 (1.13)	7.41 (0.99)	5.54 (1.81)	5.66 (1.86)			

Descriptive statistics [Mean and Standard Deviation (SD)] of the responses to each question and summary of results of the ANOVA investigating the differences between low and high numerates.

TABLE 4 | Correlation analysis for the breast cancer scenario.

	Severity	Inconvenience	Difficulty	Worry	Prior probability	Posterior probability
Severity						
Inconvenience	0.04					
Difficulty	0.003	−0.02				
Worry	0.08	0.08	0.13			
Prior probability	−0.09	0.01	0.35**	0.27**		
Posterior probability	−0.01	−0.03	0.31**	0.14	0.29**	
Recommendation	0.31**	−0.05	0.06	0.20*	0.11	0.08

***p < 0.01; *p < 0.05.*

TABLE 5 | Correlation analysis for the hypertension scenario.

	Severity	Inconvenience	Difficulty	Worry	Prior probability	Posterior probability
Severity						
Inconvenience	−0.02					
Difficulty	−0.08	−0.05				
Worry	0.23**	0.05	−0.04			
Prior probability	0.09	−0.03	0.18*	0.29**		
Posterior probability	0.10	−0.08	0.15	0.36**	0.64**	
Recommendation	0.23**	−0.13	0.02	0.29**	0.20*	0.31**

***p < 0.01; *p < 0.05.*

(β = 0.17, $p = 0.02$ and β = 0.25, $p = 0.01$, respectively) although worry closely approached it (β = 0.16, $p = 0.054$).

DISCUSSION

Ideally, making a decision implies considering the consequences of each choice as well as the probability that they may happen. However, when dealing with affect-rich situations such as deciding on medical treatments for ourselves or our loved ones, other factors seem to come into play. Our aim in this study was to investigate the role of previous beliefs and emotions in two medical situations differing in severity, i.e., in their negative consequences. Since the comprehension of probabilities is affected by the presentation format as well as by the numeracy skills of the recipient, these two variables were also controlled. A sample of university students were presented with two scenarios concerning breast cancer (the more severe disease) and hypertension (less severe) and two preventive measures that could be used to minimize their effects. Participants were required to complete a questionnaire regarding their beliefs, emotions and perception of the probabilities provided. Importantly, in order to ensure that *a priori* conceptions were measured, some of the items had to be answered before the presentation of the medical situation. The last question required participants to rate

the extent to which they would recommend the preventive measures.

As expected, participants rated breast cancer as significantly more severe and worrying than hypertension and also recommended mammograms more frequently than low-sodium diets. Indeed, when analyzing the factors influencing recommendation, only worry and severity – not probability estimations – predicted the recommendation of mammograms. Therefore, it seems that when participants had to decide on the medical situation with the worse consequences (and presumably the more affectively charged) they completely ignored the likelihood data and based their decisions on previous beliefs and current emotions. This result corroborates previous findings indicating a "probability neglect" (Sunstein, 2002) in affect-rich choices, or the existence of an "affect heuristic" (Finucane et al., 2000).

It is worth mentioning that measures of severity and worry in the breast cancer condition did not correlate (see **Table 4**), which indicates that, although both items aimed to assess emotional reactions, they were based on different sources. Indeed, only the worry rating correlated significantly with the prior probability rating meaning that emotions measured after the presentation of the medical scenario might be more influenced by the perception of the actual data contained in it. This finding reinforces our idea that studies assessing the effects of *a priori* beliefs and emotions should measure them *before* the medical scenario is presented.

In contrast, the decision to recommend a low-sodium diet was best explained by taking into consideration the posterior probability, i.e. participants' rating of the likelihood of suffering from hypertension provided a high-sodium diet was followed, and once again, severity. Therefore, our results seem to indicate that even though all medical problems are traditionally considered as affect-rich situations, probability information is not necessarily ignored; the psychological impact of probability information might depend on each particular medical condition, and more specifically, on how negatively it is perceived.

Before continuing, a specification must be made: the two scenarios differed in several aspects apart from severity, one of them being that a low-sodium diet can directly lower the chances of suffering from hypertension while mammogram only indirectly lowers the chances of having a breast cancer with worse consequences. Nevertheless, if this was the reason participants answered differently in the two scenarios, we would have expected that participants recommended the diet to a higher degree. In contrast, they were significantly more committed to recommend mammograms. The same would happen for prior probability: even if chances of suffering from hypertension are higher than those of having breast cancer, participants were more committed to recommend the preventive measure for cancer. Therefore, our results support the conclusion that differences across scenarios are due to the perceived severity of the disease described. If the medical condition is lived as severe and worrying, people do not look at likelihood or effectiveness: they simply recommend the proposed measure. In contrast, if the medical condition is not considered as severe (hypertension), they pay attention to the presented likelihoods, as shown by the significant correlation between posterior probability and willingness to recommend as well as by the fact that posterior probability was a significant predictor of the participant's decision.

Our study also addressed two factors that are known to have an effect on the difficulty of processing and understanding likelihoods: format and numeracy. As far as format is concerned, there were two reasons for its manipulation in this experiment. On the one hand, most previous studies have found that presenting probabilities with visual aids, such as icon arrays, facilitates their comprehension compared to verbal formats such as frequencies or percentages (e.g., Brase, 2009; Garcia-Retamero and Hoffrage, 2013; Tubau et al., 2018). On the other hand, other studies have stressed that icons, being a more vivid representation of the likelihood of suffering bad consequences, may have a higher emotional impact in affect-rich contexts and may increase the perceived probability (e.g., Timmermans et al., 2008). Our results provided support for both positions. First, participants considered frequencies to be harder to understand than icon arrays. Moreover, when asked to rate the posterior probability of each medical scenario (20% vs. 33%) onto the 1-to-8 scale, they rated the probabilities displayed as icons differently but provided equivalent ratings for the scenarios presented as frequencies. We consider this as further evidence that probabilities represented iconically are processed in a more fine-grained way than frequencies and are easier to manipulate and translate into context-appropriate scales. As for the effects of format on emotions, our data also supported the hypothesis that icons have a higher emotional impact, although their effects were limited to specific circumstances: information provided in the form of icon arrays was judged as more worrying than that presented as frequencies only in the most severe scenario, and only by participants with low numeracy. Therefore, we confirmed that people with low numeracy are more affected by extraneous factors (i.e. factors that do not affect objective probabilities) than their peers (Reyna et al., 2009).

Numeracy had other effects as well. As expected, less skilled individuals judged the information provided as more difficult to understand. Moreover, they judged the likelihood of suffering from breast cancer to be higher than their high-skilled peers. According to Reyna et al. (2009), uncertainty about the meaning of numerical information might lead people with low numeracy to use other criteria, such as their affective interpretation of the situation to judge probabilities or make their decisions. Given that breast cancer was considered as the most severe situation, it might have been perceived by people with low numeracy as being more likely.

Overall, the results found in this research fit well inside the fuzzy trace theory proposed by Reyna (2008). According to this author, people extract *verbatim* and *gist* representations of the information conveyed. The former are literal, precise and quantitative representations, while *gist* representations answer the question "what does the information mean to that individual?," a subjective interpretation of information that would be based on education, culture and experience, and would include the affective interpretation of this information. People prefer to operate on *gist* representations and therefore their actions might seem at odds with the objective information provided. Other approaches mentioned in the introduction also

stress the role that previous beliefs, particularly emotions, play in the decision processes: the affect heuristic, the risk-as-feelings hypothesis, or the view of affect as a spotlight stress the fact that, when deciding in affect-rich contexts, outcomes are the main consideration and their actual likelihood would be either less important, dismissed (Finucane et al., 2000; Loewenstein et al., 2001; Pachur et al., 2014), or misunderstood (e.g., by the low numeracy participants in the present study).

Before concluding, we would like to mention two possible limitations of this study that future research might seek to overcome. First, despite the differences across scenarios and the effects of format and numeracy mentioned above, in general participants provided very high ratings of the probabilities in both medical scenarios and in both formats. This was probably due to the affect involved in these medical recommendations, but it may also have been due to the scale we used. Items had to be rated on an 8-point numerical rating scale with verbal anchors (e.g., hardly severe at all–very severe). A similar 7-point rating scale had already been used in the literature (see for instance Petrova et al., 2015 or Pighin et al., 2011). Furthermore, Haase et al. (2013) concluded that verbal rating, despite showing a slightly smaller correlation with the objective measures than frequency or percentage scales (0.91 vs. 1), was a sensitive measure and better predictor of intentions and decisions than other scales. However, the fact that only eight categories were used, and that the extreme ones were marked with verbal labels, might have led participants to provide a meaningful ordinal ranking of the probabilities (*gist*) displayed in the two scenarios instead of a precise (*verbatim*) estimation (Haase et al., 2013). Therefore, even if the 8-point probability ratings properly reflected the disparity of probabilities presented in both medical scenarios, they should not be taken as a direct extrapolation from a 0-100 scale.

Second, the decision participants had to make did not concern them but a friend or relative. Nevertheless, given that one of the scenarios talked about a medical condition that mostly affects women, we wondered whether they had felt particularly involved and reacted differently from men. Unfortunately, we had not controlled for gender and most of our participants were women (115 vs. 28 men). Therefore, the following information must be interpreted with caution; however, preliminary analyses suggest that there might be gender differences. Recommendation in the breast cancer scenario was best predicted by severity in the case of women, $R^2 = 0.14$, adjusted $R^2 = 0.13$; $F(1,113) = 19.07$, $p < 0.001$; $\beta = 0.38$, $p < 0.001$, and by posterior probability in men, although the data in this case failed to reach significance, perhaps because of the small sample $R^2 = 0.11$, adjusted $R^2 = 0.07$; $F(1,26) = 3.24$, $p = 0.08$; $\beta = 0.33$, $p = 0.08$. Gender effects, though, might not be due exclusively to the medical scenario; when we ran identical analyses on the hypertension situation, we found new differences between men and women. While the likelihood to recommend the preventive measure in women was best explained by worry, $R^2 = 0.06$, adjusted $R^2 = 0.05$; $F(1,113) = 7.95$, $p = 0.006$; $\beta = 0.25$, $p = 0.006$, men's behavior was predicted by a model comprising posterior probability, $\beta = 0.59$, $p = 0.001$, severity $\beta = 0.37$, $p = 0.018$ and difficulty $\beta = -0.31$, $p = 0.049$, $F(3,24) = 7.68$, $p = 0.001$. Altogether, our current data seem to indicate that in rich-affect contexts women may pay less attention to numbers than men do, although better controlled future studies might want to confirm this point.

Summarizing, previous perception of the severity of a given medical condition modulates the use of probabilistic information for decision-making. Future efforts to ensure informed consent should not only focus on providing relevant data but may also require a reassessment of previous beliefs and emotions, and, if necessary, an attempt to correct them.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the University of Barcelona's Bioethics Commission and the protocol was approved by this commission. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

ÀC had the original idea and wrote the first draft of the manuscript. ÀC, JR-F, and ET participated in the design and reviewed the manuscript. JR-F created the questionnaire. ÀC and ET performed the statistical analysis.

FUNDING

ÀC and ET were supported by the Catalan Government (2017SGR-48). JR-F was supported by the Catalan Government (2017SGR-387) and the Spanish Government (PSI2016-80061-R, AEI/FEDER, UE). ET was also supported by the Spanish Government (PSI2017-83493-R, AEI/FEDER, UE).

REFERENCES

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Appl. Cogn. Psychol.* 23, 369–381.

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191.

Finucane, M. L., Alhakami, A., Slovic, P., and Johnson, S. M. (2000).The affect heuristic in judgments of risks and benefits. *J. Behav. Decis. Mak.* 13, 1–17.

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.

Haase, N., Renkewitz, F., and Betsch, C. (2013). The measurement of subjective probability: evaluating the sensitivity and accuracy of various scales. *Risk Anal.* 33, 1812–1828.

Harris, A. J. L., Corner, A., and Hahn, U. (2009). Estimating the probability of negative events. *Cognition* 110, 51–64.

Johnson, E. D., and Tubau, E. (2013). Words, numbers, and numeracy: diminishing individual differences in Bayesian reasoning. *Learn. Individ. Differ.* 28, 34–40.

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

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

Loewenstein, G. F.,Weber, E. U., Hsee, C. K.,and Welch, E. S. (2001). Risk as feelings. *Psychol. Bull.* 127, 267–286.

McDowell, M., and Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychol. Bull.* 143, 1273–1312.

Pachur, T., and Galesic, M. (2012). Strategy selection in risky choice: the impact of numeracy, affect, and cross-cultural differences. *J. Behav. Decis. Mak.* 26, 260–271.

Pachur, T., Hertwig, R., and Wolkewitz, R. (2014). The affect gap in risky choice: affect-rich outcomes attenuate attention to probability information. *Decision* 1, 64–78.

Peters, E., Hart, P. S., and Fraenkel, L. (2011). Informing patients: the influence of numeracy, framing, and format of side effect information on risk perceptions. *Med. Decis. Mak.* 31, 432–436.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., and Dickert, S. (2006). Numeracy and decision making. *Psychol. Sci.* 17, 407–413.

Petrova, D., Garcia-Retamero, R., Catena, A., and van der Pligt, J. (2016). To screen or not to screen: what factors influence complex screening decisions? *J. Exp. Psychol. Appl.* 22, 247–260.

Petrova, D., Garcia-Retamero, R., and Cokely, E. T. (2015). Understanding the harms and benefits of cancer screening: a model of factors that shape informed decision making. *Med. Decis. Mak.* 35, 847–858.

Petrova, D., Van der Plight, J., and Garcia-retamero, R. (2014). Feeling the numbers: on the interplay between risk, affect, and numeracy. *J. Behav. Decis. Mak.* 27, 191–199.

Pighin, S., Bonnefon, J.-F., and Savadori, L. (2011). Overcoming number numbness in prenatal risk communication. *Prenat. Diagn.* 31, 809–813.

Reyna, V. F. (2008). A theory of medical decision making and health: fuzzy-trace theory. *Med. Decis. Mak.* 28, 829–833.

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

Schwartz, L., Woloshin, S., Black, W., and Welch, G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Ann. Intern. Med.* 127, 966–972.

Slovic, P., Finucane, M. L., Peters, E., and MacGregor, D. (2004). Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. *Risk Anal.* 24, 311–322.

Sunstein, C. R. (2002). Probability neglect: emotions, worst cases, and law. *Yale Law J.* 112, 61–107.

Timmermans, D., Ockhuysen-Vermey, C., and Henneman, L. (2008). Presenting health risk information in different formats: the effect on participants cognitive and emotional evaluation and decisions. *Patient Educ. Couns.* 73, 443–447.

Tubau, E., Rodríguez-Ferreiro, J., Barberia, I., and Colomé, À. (2018). From reading numbers to seeing ratios: a benefit of icons for risk comprehension. *Psychol. Res.* doi: 10.1007/s00426-018-1041-4 [Epub ahead of print].

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

Items of the Numeracy Scale of Lipkus et al. (2001) used to test the participants' numeracy:

(1) In the BIG BUCKS LOTTERY, the chances of winning a $10.00 prize are 1%. What is your best guess about how many people would win a $10.00 prize if 1,000 people each buy a single ticket from BIG BUCKS?
(2) Imagine that we roll a fair, six-sided die 1,000 times. Out of 1,000 rolls, how many times do you think the die would come up even (2, 4, or 6)?
(3) The chance of getting a viral infection is 0.0005. Out of 10,000 people, about how many of them are expected to get infected?
(4) In the ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets of ACME PUBLISHING SWEEPSTAKES win a car?

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 July 2018
doi: 10.3389/fpsyg.2018.01136

The Reciprocal Relationships Between Escalation, Anger, and Confidence in Investment Decisions Over Time

Alexander T. Jackson[1], Satoris S. Howes[2], Edgar E. Kausel[3], Michael E. Young[4] and Megan E. Loftis[1]*

[1] Department of Psychology, Middle Tennessee State University, Murfreesboro, TN, United States, [2] Department of Management, Oregon State University, Bend, OR, United States, [3] School of Management, Pontifical Catholic University of Chile, Santiago, Chile, [4] Department of Psychological Sciences, Kansas State University, Manhattan, KS, United States

Research on escalation of commitment has predominantly been studied in the context of a single decision without consideration for the psychological consequences of escalating. This study sought to examine (a) the extent to which people escalate their commitment to a failing course of action in a sequential decision-making task, (b) confidence and anger as psychological consequences of escalation of commitment, and (c) the reciprocal relationship between escalation of commitment and confidence and anger. Participants were 110 undergraduate students who completed a series of investment decisions regarding a failing endeavor. Results revealed that although a high proportion of individuals escalate through all decisions, the extent to which they escalated decreased with each decision as they were less willing to invest money in the project. Furthermore, as participants escalated, confidence in one's decision decreased and anger increased. Lastly, the analyses revealed that the relationship between escalation and confidence is reciprocal. Escalation was negatively associated with confidence, and confidence predicted escalation in the subsequent decision. These results highlight the importance of considering both the determinants and psychological consequences of escalation of commitment.

Keywords: escalation of commitment, sequential decision making, confidence, anger, judgments

OPEN ACCESS

Edited by:
Gorka Navarrete,
Adolfo Ibáñez University, Chile

Reviewed by:
Kin Fai Ellick Wong,
Hong Kong University of Science and Technology, Hong Kong
Richard S. John,
University of Southern California, United States

****Correspondence:***
Alexander T. Jackson
alexander.jackson@mtsu.edu

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

Received: *27 January 2018*
Accepted: *14 June 2018*
Published: *05 July 2018*

Citation:
Jackson AT, Howes SS, Kausel EE, Young ME and Loftis ME (2018) The Reciprocal Relationships Between Escalation, Anger, and Confidence in Investment Decisions Over Time. Front. Psychol. 9:1136. doi: 10.3389/fpsyg.2018.01136

INTRODUCTION

> "You gotta know when to hold 'em, know when to fold 'em, know when to walk away, know when to run."
>
> -Kenny Rogers, The Gambler

The above lyric epitomizes the psychological bias of escalation of commitment. Escalation of commitment refers to the tendency to invest additional resources into an ongoing effort, when doing so is no longer rational (Staw, 1976; Sleesman et al., 2012). Continuing to invest funds into a failing project can result in large financial and productivity losses (e.g., Arkes and Blumer, 1985; Ross and Staw, 1993; Schmidt and Calantone, 2002; Drummond, 2014). Furthermore, research has demonstrated that raters will positively bias their assessments of those employees they were responsible for hiring, thereby demonstrating that escalation of commitment is an issue

within personnel selection and assessment decision contexts (Schoorman, 1988). Consequently, considerable attention has been directed toward understanding the factors that influence the likelihood of escalating commitment to a failing course of action. For example, Sleesman et al. (2012) reviewed over 160 studies and found strong evidence that previous resource expenditures, such as money and time, are associated with higher levels of escalation.

Despite the abundance of research on escalation of commitment, the majority has examined a single decision to escalate, despite organizational decision makers often facing repeated decisions about underperforming projects or personnel. Some research has shown a clear trend that individuals initially escalate their commitment to a failing course of action, especially if they were responsible for the initial decision (Staw, 1976). However, the results regarding what happens *after* the initial escalation are mixed. Staw and Fox (1977) found that the relationship between time and escalation resembled a U-shaped function. Immediately following the initial investment decision, participants invested additional funds into a failing venture, thereby exhibiting a strong escalation bias. After receiving additional negative feedback about the venture, participants tended to de-escalate by investing less money in the next decision. When presented with a third decision, however, participants invested significantly more money than the second decision. In contrast, McCain (1986) demonstrated that over the course of 10 financial decisions, individuals tended to escalate their commitment with the first decision after failure, and de-escalate with subsequent decisions.

In addition, the research examining escalation of commitment has predominantly focused on the antecedents of escalating, lacking an evaluation of what happens after escalation has occurred (Schoorman and Holahan, 1996; Sleesman et al., 2012). Specifically, most studies have focused primarily on understanding the influence of project characteristics and psychological factors on the decision to escalate (Sleesman et al., 2012). Little attention, however, has been paid to other aspects of escalation situations, such as the psychological consequences of escalating. These consequences are important, as consequences of escalating might impact future decisions. For instance, while recent research has shown that (over)confidence indeed positively predicts escalation behavior (e.g., Ronay et al., 2017) what is currently unknown is how escalation behavior impacts subsequent confidence. We attempt to fill this gap in the literature by examining two psychological consequences: confidence and anger.

Another area of research that warrants additional focus is whether participants would choose to continue with or abandon the failing project and the repercussions of this decision to abandon. Staw and Fox (1977) allowed their participants to invest \$0 in a failing project, whereas McCain (1986) permitted their participants to choose to quit the simulation after the third decision. The current study seeks to extend the findings of Staw and Fox (1977) and McCain (1986) by allowing participants to abandon the project after the first decision and throughout the subsequent decisions. Though Staw and Fox (1977) allowed participants to invest \$0 from the beginning, the psychological difference between investing \$0 and abandoning the project could be impactful. For example, investing \$0 in the project likely leads a participant to believe that the project will continue, just without the additional funds requested. Participants may even feel as though they could invest additional funds into the project later, if the project started to become more promising. In contrast, abandoning the project means that the project will end and the participant is completely giving up hope that the project may be successful. Proceeding in this manner allows us to not only evaluate participant decisions, but also judgments. According to Bonaccio and Dalal (2006), a decision is a choice – such as a participant's choice to invest additional funds or abandon the project. In contrast, a judgment is a quantitative value – such as the amount of funds actually invested, which may reflect their confidence in the decision to invest.

In summary, there are three purposes of the present study. First, we aim to examine how escalation of commitment unfolds over a series of decisions. Second, we seek to evaluate the degree to which participants escalate to a failing venture by using both their decision to continue funding and their judgment regarding how much to invest. Lastly, whereas previous studies have focused on the determinants of escalation, we evaluate the psychological consequences of anger and confidence, as well as their reciprocal nature, on future escalations in a sequential decision-making context.

Escalation of Commitment

By definition, escalation of commitment – the decision to invest resources toward an endeavor that one knows is failing – is irrational. When faced with the decision to continue with a course of action or abandon a failing endeavor, the rational choice would be to ignore the time, energy, and resources already invested (i.e., the sunk costs) and abandon the project. Nevertheless, people *do* heavily consider sunk costs and often make the irrational choice of escalating their commitment to failing endeavors (Conlon and Garland, 1993).

The act of escalating itself is not necessarily irrational. Sunk cost effects do not only depend on escalation behavior; they also require an examination of the rationale for such behavior. There may be multiple possible forces promoting escalation behavior, such as overconfidence, the presence of sunk costs, the social and reputational damage of admitting failure, organizational or political barriers, and the need for self-justification (Drummond, 2014). At the same time, there may be multiple possible factors placing pressure on the decision maker to abandon the endeavor, such as one's own loss aversion, being perceived as wastefully expending resources that could be spent on other opportunities, and the political pressures associated with publicly stated limits or stopping points (Drummond, 2014). Therefore, one must consider the reasons for escalating to determine whether sunk cost effects are occurring.[1]

One of the most promising explanations for initial escalation behavior involves self-justification theory (Brockner, 1992). Opting to terminate a failing project or dismiss an ineffective employee would create cognitive dissonance for individuals, as their decision would be counter to their initial belief

[1]We would like to thank a reviewer for highlighting this point.

that the project or employee would be a success. As such, individuals who were responsible for the initial decision attempt to justify their actions in order to reduce the cognitive dissonance they are experiencing (Sleesman et al., 2012). By rationalizing their behavior and electing to continue with the failing project/employee (hoping things will ultimately turn around and initial beliefs will be justified), individuals are able to protect their ego (Zhang and Baumeister, 2006) and reduce any further cognitive dissonance. For instance, when faced with an escalation decision, concerns about one's own reputation may be heightened (Zhang and Baumeister, 2006; Sleesman et al., 2012). The heightened reputational concerns activate the need to justify one's decisions. If a previously decided course of action is now failing, the decision maker may feel that his or her reputation is going to be harmed because he or she is not succeeding. This may then lead to the need to justify one's previous actions by staying the course with the hopes of eventual success. Indeed, ego threat (i.e., reputational threat) has been shown to be one of the strongest predictors of escalation (Sleesman et al., 2012).

In support of the self-justification theory, Arkes and Blumer (1985) demonstrated that the money already invested in a particular venture leads people to experience pressure to justify their actions to themselves. In an effort to avoid appearing wasteful, individuals opt to continue a course of action with the hope that the venture will ultimately be successful. In a similar Wong and Kwong (2007) found that when anticipated regret for abandoning a project early was high, individuals were more likely to escalate their commitment. As such, anticipated regret appears to serve as self-justification mechanism for escalation. These are only two of many determinants possible mechanisms for self-justifying escalation behavior. For instance, in their meta-analysis, Sleesman et al. (2012) found support for a number of psychological determinants of escalation of commitment: sunk costs, time investment, personal experience or expertise, self-efficacy/confidence, responsibility for the initial decision, ego threat, and anticipated regret. Each of these determinants relies on self-justification theory to explain why people escalate. Accordingly, Sleesman et al. (2012) argued that self-justification theory has merit as a central theory explaining *why* individuals choose to escalate their commitment toward failing endeavors.

Escalation in Sequential Decisions

One of the key characteristics of decision making in applied settings is that many dilemmas, including escalation decisions, are not resolved after a single decision. Even the dilemma of whether to continue with or abandon a failing endeavor may have multiple decision points. For example, after making an initial investment decision, an employee may receive negative feedback that triggers self-justification processes and the decision to invest additional resources. After those resources are used, the project may still be failing, and additional decisions may be required to continue the project. Initial decisions, the associated outcomes of those initial decisions, subsequent decisions, and outcomes all unfold over time. Thus, it is important to examine the extent to which people are willing to repeatedly escalate.

As noted previously, self-justification theory argues that individuals feel compelled to justify their previous behavior, which in turn leads to escalation. In the case of a sequential decision-making situation, individuals may repeatedly receive information that their previous decisions were poor, continuously reviving the cognitive dissonance experienced (Draycott and Dabbs, 1998). Accordingly, individuals would be expected to engage in self-justification processes and continue escalating to alleviate the renewed dissonance experienced at each decision point. However, over time, it may become more difficult to rationalize one's actions and reduce the cognitive dissonance. Instead, the only feasible option becomes reducing the investment in a project, and ultimately discontinuing one's commitment despite sunk costs.

In addition, cognitive dissonance may lead individuals to experience psychological discomfort (Festinger, 1957; Elliot and Devine, 1994), negative affect (Harmon-Jones, 2000), and physiological arousal (Elkin and Leippe, 1986). However, even when individuals engage in strategies to alleviate these negative sequelae, they may not experience a decrease in arousal (Elkin and Leippe, 1986). Therefore, although individuals engage in self-justification processes, it is likely that they do not experience a decrease in psychological discomfort. As such, we expect the discomfort experienced from repeatedly learning that one's decisions are not leading to success will result in increased anger at the decision, and reduced confidence in one's ability to make good decisions. Support for the above notion comes from the feedback literature. Both Tata (2002) and Belschak and Den Hartog (2009) documented the effect of performance feedback from managers on employees' affect and found that when performance feedback was negative, employees tended to experience more anger. In line with this, we predict that as decisions unfold over multiple decision points and negative feedback continues, individuals are less likely to escalate their commitment (both in terms of commitment to the project and monetary investment toward the project) and are likely to become increasingly angry and decreasingly confident in their decisions. This leads to the following hypotheses:

Hypothesis 1: Escalation of commitment toward a failing course of action will decrease over time (multiple decision points).
Hypothesis 2: The degree to which individuals escalate (invest money) toward a failing course of action will decrease over time (multiple decision points).
Hypothesis 3: Continued escalation over multiple decisions will lead to increased anger.
Hypothesis 4: Continued escalation over multiple decisions will lead to decreased confidence.

Further, these relationships may be reciprocal in nature. Supporting this idea, Van Overwalle and Jordens (2002) argued that individuals rely on information about their affective experiences when making judgments. According to affective events theory, work environment features influence work events, work events lead to affective reactions, and affective reactions ultimately lead to affect driven behaviors (Weiss and Cropanzano, 1996). Affective events theory can be used to explain the determinants of escalation by examining the

work environment features (i.e., features of the decision), the psychological outcomes of work events (i.e., electing to escalate), and future behaviors. Specifically, when an individual receives negative feedback about a course of action and elects to continue with the course of action, he or she would likely experience the hypothesized changes in confidence and anger. The hypothesized decrease in confidence would then lead an individual to actually abandon the failing course of action sooner. In other words, because an individual becomes less confident in their decision-making capabilities as a result of escalating, he or she would be less likely to continue escalating, resulting in a downward spiral of confidence and escalation. Similarly, the hypothesized increase in anger resulting from escalating likely leads an individual to abandon the endeavor sooner, rather than later (**Figure 1** displays this conceptual model). Indeed, Strough et al. (2016) found that negative affect is positively related to willingness to cancel a failing plan. This leads to the following hypotheses:

Hypothesis 5: The lagged effect of anger negatively predicts escalation in the next decision.

Hypothesis 6: The lagged effect of confidence positively predicts escalation in the next decision.

MATERIALS AND METHODS

Participants

Participants were 110 (32 males, 78 females) undergraduate students recruited from a large Midwestern university. The sample was primarily White, non-Hispanic (84%), with a mean age of 20 (SD = 3). Approximately 52% of the sample was employed at least part-time.

Procedures

The study was conducted individually in a laboratory setting. The decision task was based on the "blank radar plane" case originally presented by Arkes and Blumer (1985) and widely used to study escalation of commitment (Conlon and Garland, 1993; Moon, 2001a,b; Wong et al., 2006). For this task, participants are presented with a vignette in which they are asked to assume the role of the Vice President of Operations for a mid-sized high-technology manufacturing firm. Because we were interested in examining how escalation of commitment occurs over time, we created additional decision vignettes that followed the first decision (each of the decision vignettes are presented in **Appendix A**). All information was presented in printed text format.

Participants were presented with the first scenario and asked to make the initial investment decision. Specifically, they were asked, "Between 5 million dollars and 10 million dollars, how much money would you like to invest in the project?" For each subsequent decision, participants were presented with additional information indicating that the project was still not complete and needed additional funds (see **Appendix A**). Participants were then asked whether they wanted to authorize more funds to continue the project or abandon the project. The specific instructions stated, "The decision you face now is to either abandon the project or authorize more funding to continue this radar-scrambling project." Thus, the escalation variable was dichotomous (continue the project vs. abandon the project). Additionally, if participants chose to authorize more funding, they were asked how much money they wished to authorize, based on the information from the vignette. The amount of money participants could choose to authorize differed throughout the five exercises and was based on the current state of the project in order to increase the fidelity of the task (see **Appendix A**). See **Table 1** for the differing funding available at each decision point. If individuals chose to pursue the failing project at all decision points, they would be asked to make a total of five decisions regarding the funds to be authorized and four decisions regarding whether to continue with the project (the escalation vignettes are available upon request from the authors). Immediately following each decision point, participants were asked to indicate how confident they were in their decision and to complete a brief measure of emotions. Performance in the task was not incentivized. The study ended after a participant continued through all five decisions or abandoned the project.

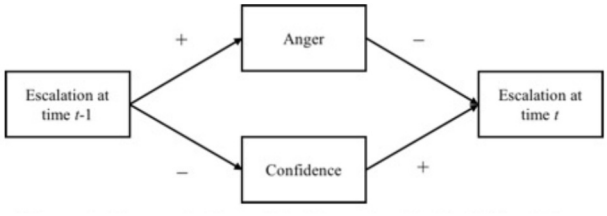


FIGURE 1 | Conceptual model of hypothesized relationships.

Measures

Confidence

Confidence was measured using two items adapted from Greer and Stephens (2001) confidence scale. After each decision point, participants indicated how confident they felt in their decision using a 1 (*Low Confidence*) to 7 (*High Confidence*) scale. An example item stated, "How confident are you in your resource investment decision?" Responses to the two items were averaged to create a composite confidence score for each decision point. The average coefficient alpha for the confidence scale across the five decision points was 0.92.

Anger

To assess how anger changes over time, we used the four anger-related items (i.e., Angry, Furious, Mad, and Frustrated) from the

TABLE 1 | Funding available at each decision point.

Decision 1	$5 million – $10 million
Decision 2	$3 million – $6 million
Decision 3	$2 million – $5 million
Decision 4	$4 million – $7 million
Decision 5	$1 million – $4 million

State Trait Anger Expression Inventory (Spielberger et al., 1988). After each decision, participants were asked, "When thinking about the events that led to you having to make this decision, how are you feeling right now?" Participants responded using a 1 (*Not at all like me*) to 5 (*Very much like me*) scale. In order to minimize any priming effects for the subsequent decision, the anger items were interspersed with an equal number of positively valenced items. These positive items were not examined because they were used to minimize demand characteristics. The average coefficient alpha for the state anger scale across the five decision points was 0.83.

Additional Measures Not Included in the Analyses

This study was conducted as part of a larger study examining escalation of commitment in general. As a part of this larger study, data on additional variables were also collected during the data collection. These variables include a variety of personality traits, including: generalized self-efficacy, grit, regulatory focus, goal orientation, narcissism, psychopathy, Machiavellianism, trait affectivity, trait regret, guilt proneness, and the big five. Additionally, participant's state level of pride was measured. However, the focus of this study was exclusively on the state changes in confidence and anger as people escalate. As such, personality characteristics and pride were excluded from the analyses.

RESULTS

Correlations between all variables of interest are displayed in **Table 2**. First, we examined how long individuals were willing to continue with a failing course of action. A repeated measures logistic regression was conducted using the generalized linear mixed-effects modeling package in *R* (Bates et al., 2014). A logistic regression was selected because escalation was measured as a dichotomous variable (authorize funds vs. abandon project). Because the current analysis assessed escalation over time, decision number (the effect of time) was entered as a continuous fixed effect, participant was entered as a random intercept effect[2], and escalation was entered as the criterion. The results of the analysis revealed that decision number was significantly negatively associated with escalation of commitment, $B = -0.92$, $z = -4.02$, $p < 0.01$, 95% CI [−1.36, −0.47]. In other words, as the project progressed and negative feedback continued, participants became less willing to continue with the project, supporting Hypothesis 1. As shown in **Figure 2**, the predicted proportion of individuals escalating their commitment after the initial investment decision is approximately 0.98, yet over time, this proportion significantly decreased to 0.68 at the fourth decision after the initial investment decision.

Next, we examined how the amount of money individuals are willing to invest changes over time. Participants were given a range of funds to invest at each decision point. Because different ranges of available funds were used at each decision point based on the scenario, the amount of money each person invested was transformed into the proportion of money invested out of the total amount available at that decision point. For individuals who selected the maximum or minimum values of the range of funds available, their proportions were 1.00 and 0.00, respectively. Data that are scored from 0.00 to 1.00, such as proportions, often accumulate heavier at the 0.00 and 1.00 values than the values of a normal distribution, necessitating a logit transformation to extend the tails of the distribution (Cohen et al., 2003). With a logit transformation, data equaling 0.00 and 1.00 become negative and positive infinity, respectively, in the transformed space, which necessitates adding 0.05 to any 0.00 values and subtracting 0.05 from any 1.00 values. Thus, for each of the proportions that were 1.00 or 0.00, 0.05 was subtracted or added, such that adjusted scores ranged from 0.05 to 0.95. These adjusted scores remained the highest and lowest values on the scale. The proportion of funds invested variable was then logit transformed.

A linear mixed effects regression was then conducted using the linear mixed-effects modeling package in *R* (Bates et al., 2014). As in the first analysis, because we are examining changes over time, decision number (the effect of time) was entered as a fixed effect, participant was entered as a random intercept

[2]The nature of the task does not allow the time slope to vary across subjects because it produces an identical slope for each participant. When a participant chooses to abandon the task, the value for escalation becomes 0. Accordingly, we did not allow the slopes to vary for participants.

TABLE 2 | Correlations among variables.

Variable	1	2	3	4
(1) Decision number	–	–	–	–
(2) Escalation	**−0.27**	–	–	–
(3) Proportion invested	**−0.33**	**0.11**	–	–
(4) Confidence	**−0.17**	0.07	**0.18**	–
(5) Anger	**0.29**	**−0.12**	−0.05	**−0.33**

Bolded values are significant at $p < 0.05$.

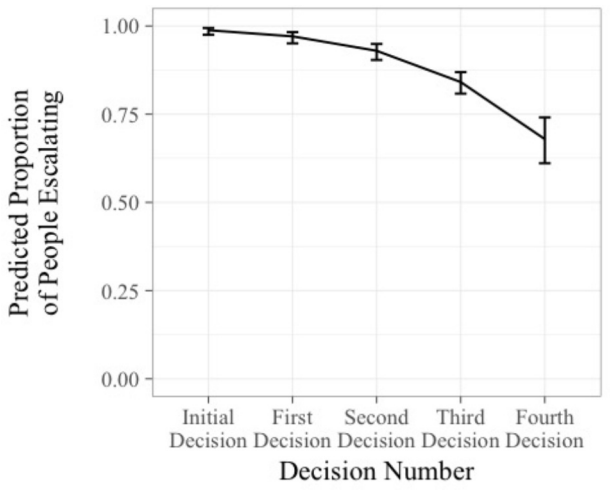


FIGURE 2 | Predicted probability of escalating commitment over time. Error bars represent ±1 standard error.

effect, and the proportion of funds invested was entered as the criterion. We also evaluated models that included decision number (the time slope) as a random slope effect. However, because the model excluding decision number as a random effect resulted in a better fitting model (i.e., lower BIC), we report and interpret the results of model with decision number entered as only a fixed effect. Results revealed that decision number was negatively associated with the proportion of funds people were willing to invest, $B = -0.68$, $t(428) = -8.44$, $p < 0.05$, 95% CI [−0.84, −0.52]. Thus, as the course of action continues to fail, participants became increasingly less willing to invest money over time, supporting Hypothesis 2. As shown in **Table 3** and **Figure 3**, the predicted proportion of available funds initially invested was 0.26 ($6.3 Million when choosing from the $5 to $10 million range). However, as participants continued to escalate their commitment over time, the predicted proportion of funds approached the minimum of the offered range. In other words, at the initial investment decision, participants were willing to invest more than the first quartile of the available funds. However, as they continued to escalate their commitment, willingness to invest dropped substantially. At the final decision, individuals invested near the minimum of the available funds.

Hypotheses 3 and 4 stated that individuals would experience increased anger and decreased confidence, respectively, as they continue to escalate their commitment toward the failing project. Separate analyses were conducted with anger and confidence as the criterion in each analysis. According to Bonaccio and Dalal (2006), a decision is a choice – such as a participant's choice to invest additional funds or abandon the project. In contrast, a judgment is a quantitative value – such as the amount of funds actually invested, which may reflect their confidence in the decision to invest. We operationalized escalation both as a decision (continue funding) and a judgment (proportion invested). Therefore, we conducted two separate linear mixed effects regressions. In both analyses, decision number (the effect of time) was entered as a fixed effect and a random slope effect, and participant was entered as a random intercept effect. The operationalization of escalation was entered as a fixed effect. We also examined models with the operationalization of escalation entered as a random slope effect. However, in all of the models examined, the best fitting model (i.e., lowest AIC and BIC) did not include escalation as a random slope effect. Therefore, we report the results of the models with the operationalization of escalation entered as a fixed effect only.

TABLE 3 | Proportion of funds invested over time.

	Available funds	Relative proportion invested	Average amount invested
Decision 1	$5 million – $10 million	0.26	$6.3 million
Decision 2	$3 million – $6 million	0.15	$3.45 million
Decision 3	$2 million – $5 million	0.08	$2.24 million
Decision 4	$4 million – $7 million	0.04	$4.12 million
Decision 5	$1 million – $4 million	0.02	$1.06 million

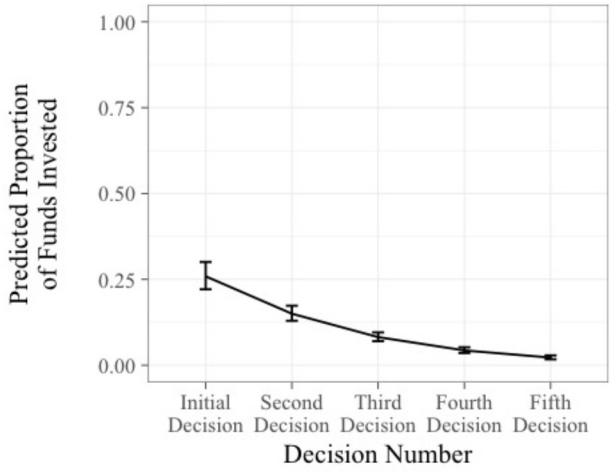


FIGURE 3 | Predicted proportion of funds invested over time. Error bars represent ±1 standard error.

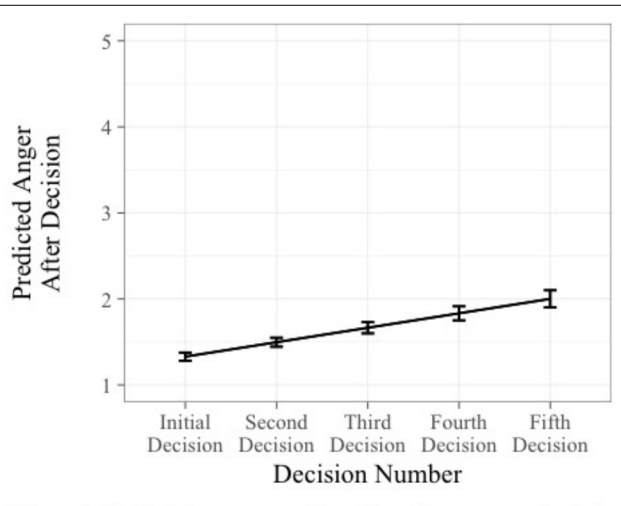


FIGURE 4 | Predicted anger over time. Error bars represent ±1 standard error.

The first set of analyses examined anger as the criterion. The first regression operationalized escalation as a decision and revealed that the decision to escalate was not a significant predictor of anger, $B = -0.11$, $t(481) = -1.19$, $p > 0.05$, 95% CI [−0.28, 0.07]. However, decision number was positively associated with anger, $B = 0.16$, $t(481) = 6.50$, $p < 0.05$, 95% CI [0.11, 0.20]. Thus, with each decision people became increasingly angry (see **Figure 4**). The second regression operationalized escalation as a judgment by using the proportion of funds invested variable. This second regression revealed a similar pattern. The proportion of funds invested did not significantly predict anger, $B = -0.02$, $t(428) = -1.43$, $p > 0.05$, 95% CI [−0.04, 0.01], but decision number was positively associated with anger. Therefore, Hypothesis 3 was not supported.

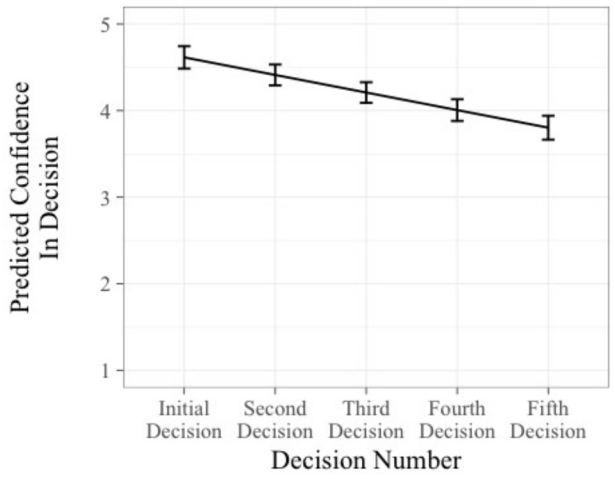


FIGURE 5 | Predicted confidence over time. Error bars represent ±1 standard error.

The second set of analyses examined confidence as the criterion. The first regression operationalized escalation as a decision and showed that the decision to escalate did not significantly predict confidence in the decision, $B = -0.23$, $t(476) = 1.48$, $p > 0.05$, 95% CI [−0.54, 0.08]. However, decision number significantly predicted confidence, $B = -0.23$, $t(476) = -5.90$, $p < 0.05$, 95% CI [−0.31, −0.16]. This suggests that as individuals continue with a failing course of action, they become decreasingly confident with each decision that they are making (see **Figure 5**). The second regression, operationalizing escalation as a judgment, revealed that the proportion of funds invested *did* significantly predict confidence, $B = 0.06$, $t(423) = 2.97$, $p < 0.05$, 95% CI [0.02, 0.10]. Furthermore, decision number was negatively associated with confidence, $B = -0.19$, $t(423) = 4.67$, $p < 0.05$, 95% CI [−0.27, −0.11]. Thus, as people invested more money in each decision, they felt more confident in the decision. However, on average confidence still decreased with each decision. Thus, Hypothesis 4 was partially supported.

Lagged Effects of Anger and Confidence

In order to test Hypotheses 5 and 6, the lagged effects of anger and confidence on escalation behavior were examined. Specifically, we examined the effect of anger and confidence at time *t-1* on the decision to escalate at time *t* in separate analyses. The analyses were conducted using the linear mixed effects modeling and generalized linear mixed effects modeling packages in *R*. Additionally, like in the analyses testing Hypotheses 3 and 4, separate analyses were conducted for the separate operationalizations of escalation.

Hypothesis 5 stated that the lagged effect of anger would be negatively associated with escalation in the next decision. To test this, the operationalization of escalation at time *t* was predicted by anger at time *t–1* and decision number at time *t*. Specifically, anger at time *t*–1 and decision number were entered as fixed effects, and participant was entered as a random intercept effect. The lagged effect of anger at time *t-1* did not significantly predict whether one chose to escalate at time *t*, $B < 0.01$, $z = 0.02$, $p > 0.05$, 95% CI [−0.46, 0.47]. However, decision number significantly predicted the choice to escalate at time *t*, $B = -0.65$, $z = -2.74$, $p < 0.01$, 95% CI [−1.11, −0.18]. A similar pattern emerged when examining escalation as the amount of funds invested. The lagged effect of anger at time *t−1* did not significantly predict the funds one invested at time *t*, $B = 0.39$, $t(372) = -1.80$, $p > 0.05$, 95% CI [−0.04, 0.81]. However, decision number significantly predicted the funds one invested at time *t*, $B = -0.32$, $t(372) = -2.74$, $p < 0.05$, 95% CI [−0.55, −0.09]. It appears that although people get increasingly angry with each decision, the increased anger does not influence the choice to escalate or the funds one chooses to invest in subsequent decisions. Therefore, Hypothesis 5 was not supported.

Hypothesis 6 stated that the lagged effect of confidence would be positively related to escalation in the next decision. To test this, the decision to escalate at time *t* was predicted by confidence at time *t−1* and decision number at time *t*. Specifically, confidence at time *t−1* and decision number were entered as fixed effects, and participant was entered as a random intercept effect. The lagged effect of confidence at time *t−1 did* significantly predict whether one chose to escalate at time *t*, $B = 0.27$, $z = 2.00$, $p < 0.05$, 95% CI [0.01, 0.54]. Additionally, decision number significantly predicted the choice to escalate at time *t*, $B = -0.62$, $z = -2.61$, $p < 0.01$, 95% CI [−1.08, −0.15]. In contrast, the lagged effect of confidence at time *t−1 did not* significantly predict the funds invested at time *t*, $B = 0.09$, $t(316) = 0.81$, $p > 0.05$, 95% CI [−0.13, 0.32]. Although people become decreasingly confident with each decision, the decreased confidence does not influence the choice to escalate in the next decision. However, individuals who were more confident after each decision invested slightly more of the available funds in the next decision. Therefore, Hypothesis 6 was partially supported.

DISCUSSION

The purpose of this study was threefold. First, we sought to examine the extent to which people continue escalating their commitment in a sequential decision-making situation. The results of this study revealed that when faced with sequential escalation decisions, fewer and fewer individuals escalate their commitment over time. However, as can be seen in **Figure 2**, a substantial 68% of participants never abandoned the project and escalated through all five decisions. Thus, it seems that when faced with an escalation of commitment dilemma, individuals are highly likely to continue pursuing the failing course of action over time. Furthermore, the results demonstrated that the degree to which individuals tended to escalate decreased over time, as they invested decreasing proportions of the available funds with each decision. These findings demonstrate that while individuals do continue with a failing course of action, they become less willing to invest money in the endeavor with each decision. This may indicate that people feel a decreasing amount of confidence in their ability to turn the project around. Alternatively, the adaptive learning strategies model proposed by

Wong and Kwong (2017) may be used to explain our findings. Specifically, they argue that the probability that one will escalate his or her commitment is a function of learning at the strategy level, not the individual decision level. If a specific strategy (e.g., escalation) is reinforced, the individual should continue that strategy. However, if the strategy is not reinforced, the individual should change strategies. In our present study, the escalation was not reinforced. Indeed, the escalation strategy in our study was punished with repeated negative feedback. As such, we see the increasing degree of abandonment with later decisions, even though 68% of participants never abandoned the project.

Second, we sought to examine two psychological consequences (anger and confidence) of escalating one's commitment to a failing course of action. In the current study, we examined anger and confidence as psychological outcomes and operationalized escalation as a decision (continue to escalate) or a judgment (proportion of funds invested). According to self-justification theory, electing to abandon or reduce the amount of funding to a failing project would create cognitive dissonance for individuals because this would contradict their initial belief that the project would succeed. This cognitive dissonance leads to psychological discomfort, such as anger (Harmon-Jones, 2000). As such, we argued that individuals may actually experience an increase in anger and a decrease in confidence with each subsequent decision or judgment as this revitalizes the initial cognitive dissonance and psychological discomfort. While our results showed that neither the decision to escalate nor the judgment regarding the funds to invest significantly predicted anger, individuals did become increasingly angry with each escalation decision. This may be the result of repeatedly receiving negative feedback because their decisions kept resulting in negative outcomes and the project was not succeeding. Indeed, Belschak and Den Hartog (2009) found that receiving negative feedback led to increased negative affect.

Regarding confidence, we found that the decision to escalate did not significantly predict confidence. However, the judgment regarding the funds to invest significantly predicted confidence, such that as people invested a higher proportion of the funds they felt more confident. Interestingly, as people continued with each decision, they became decreasingly confident. Again, this is likely the result of repeatedly receiving negative feedback. As previously noted, people decreased the proportion of available funds they invested over time. We argue that the decrease in investment with each decision may be a reflection of the decrease in confidence.

The third and final purpose of this study was to explore the lagged effects of confidence and anger on one's decision to escalate in future decisions and one's judgment of the amount of money to invest in future decisions. We found that the lagged effect of anger did not influence one's future decision to escalate or the proportion of funds invested in the next decision. This suggests that while people continue to get angry, perhaps as a result of the negative feedback or lack of success, the increasing anger may simply be reactionary and does not influence their subsequent decisions. In other words, people get angry about their past decisions but do not let it affect their future decisions. Our results are similar to the findings of Tsai and Young (2010). Specifically, they found that although individuals induced to feel angry were significantly more likely to escalate their commitment, they were not significantly more likely to escalate than individuals induced with a neutral emotional prime.

Although our hypotheses regarding anger were not supported, they are in line with the Appraisal Tendency Framework (ATF). According to the ATF, emotions, such as anger, carry motivational properties that influence judgments and decisions (Han et al., 2007). Furthermore, these motivational properties influence the contents of a person's thoughts, such that anger may lead individuals to blame others for negative events and believe he or she can still have positively influence the situation (Lerner and Tiedens, 2006). This is particularly important in escalation situations because the decision maker may experience anger when he or she receives information about the project's lack of success. Accordingly, this anger may lead the decision maker to blame others for the negative events while still hoping that the decision maker could turn the project around. Indeed, one reason people continue escalating their commitment to a failing course of action is because they think they can turn the failing project around (Sleesman et al., 2012).

The story for confidence is slightly different. Our results demonstrated that the lagged effect of confidence *did* influence one's future decision to escalate. However, the lagged effect of confidence did not influence the proportion of funds invested in one's future decision. Therefore, it appears that those who had higher confidence were more likely to continue escalating but were no more likely to invest more money into the project. This is somewhat similar to the findings of Ronay et al. (2017) who demonstrated that overconfidence predicts escalation behavior in public but not private settings. Thus, the reciprocal relationship between escalation of commitment and confidence is one that is less straightforward than the reciprocal relationship between escalation of commitment and anger.

Implications

These results have important implications for practitioners. At a basic level, this study provides good and bad news regarding escalation of commitment over time. First, this study shows that while not everyone continually persists with a failing course of action, a substantial proportion of individuals in the current study (68%) never abandoned the project. Though this finding is noteworthy, one must recognize that the 68 percent may be an overestimation, as participants were not investing their own money and may have felt they could make riskier decisions. For instance, people tend to make riskier financial decisions when the money used is not their own (Chakravarty et al., 2011). However, while this value may be an overestimation, it is nevertheless noteworthy especially when considering the notion that a manager making investment decisions for a company in many circumstances is not using his or her own money.

Limitations and Future Directions

One of the primary limitations of this study is the sample we used. Notably, our sample consisted of undergraduate students completing an artificial task. Therefore, our sample may not be representative of real organizational decision makers. It is hard to say students, who are not investing their own money and from whom there are no true consequences to escalating, are comparable to a manager who is investing company resources with their job on the line. It is plausible that a manager would exhibit higher concern for the failure of the project. However, the focus of this study was on basic decision-making processes and the psychological outcomes of escalating. Had the consequences of decisions been real (i.e., actual loss of money and resources) the results may have differed. Nevertheless, we took steps to increase the fidelity of the situation, and participants appeared to take the task seriously (i.e., not arbitrarily quitting or continuing with the study). Future research should replicate the results of the study using a sample of managerial decision makers. An additional limitation of this study is that we only examined one continuously failing project. In reality, projects may have periods of success in addition to setbacks. Future research should investigate how participants would react if they are given a glimmer of hope for the project to recover, only to have it ultimately fail in the end.

An additional goal for future researchers is to examine the individual differences, such as guilt, shame, or pride, that lead some people to continually escalate and others to abandon a failing course of action. The results of such research could assist organizations in selecting individuals who are less likely to waste organizational resources. For example, future researchers should examine the role of the dark triad (narcissism, psychopathy, and Machiavellianism; Paulhus and Williams, 2002) on escalation of commitment over time. Given the destructive nature of these traits and their relations with job performance and counterproductive work behaviors (O'Boyle et al., 2012), one would expect that individuals high on these traits would not feel guilty about wasting an organization's resources. Therefore, a positive relationship may be expected between the dark triad and escalation of commitment over time. In addition, future research should seek to try to increase the ecological validity in decision making tasks in order to decrease the chance participants are taking risks simply because there are no consequences.

Our results also demonstrate that whereas people do escalate, they become less confident in their decision and ultimately invest decreasing proportions of available funds over time. This suggests that individuals may actually recognize that they are making irrational decisions by escalating. In accordance with self-justification theory, individuals may justify their escalation actions by investing fewer and fewer resources over time. The good news for organizations is that while people may tend to make irrational decisions, they waste proportionally fewer organizational resources over time. That said, decision makers should be afforded the opportunity to abandon failing endeavors, without the possibility of negative consequences, as people may actually recognize the irrationality of the decision to escalate and may be more inclined to abandon a failing course of action if there are not possible negatives consequences associated with abandoning.

CONCLUSION

Our study contributes to the escalation literature in two meaningful ways. First, we demonstrated that escalation of commitment occurs over time, and that over half of the participants escalated through all five decisions. Second, we examined the impact of escalating on psychological consequences, such as anger and confidence. We found that with each decision individuals become angrier and less confident with their decisions. Additionally, the choice to escalate negatively impacts confidence, which then positively predicts the choice to escalate in the next decision.

Evidence of the effects found in the current study is present in well-known examples of escalation of commitment, such as Tesco's withdrawal from United States markets. After nearly 6 years of continued efforts, Tesco, one of Britain's largest supermarket companies, eventually chose to abandon their failing attempt to enter the United States supermarket industry (Werdigier, 2013). Some estimates of the cost for Tesco to withdraw from United States markets were as high as £1.5 billion on top of the $22 million each month the company was losing on its Fresh and Easy supermarket chain (Butler, 2013). In March 2008, Tesco began slowing its expansion of stores in the United States and the long-term viability of the project was increasingly in question (Best, 2012). Thus, confidence waned over time and investments became increasingly smaller. In the case of Tesco, those tasked with making optimal decisions often irrationally escalated their commitment, and such irrational decision making clearly came at a great cost.

ETHICS STATEMENT

Rick Scheidt, Kansas State University IRB Committee Chair (785) 532-1483 rscheidt@ksu.edu. Prior to participating in this research, participants provided written informed consent. Participants also had an opportunity to ask any questions regarding the study prior to signing the informed consent. After providing written informed consent, participants were presented with the decision-making tasks in this research.

AUTHOR CONTRIBUTIONS

AJ, SH, and EK conceptualization. AJ and SH data collection. AJ and MY data analysis. AJ, SH, EK, MY, and ML manuscript preparation and review.

FUNDING

Financial support from FONDECYT, under Grant Regular #1171894, is gratefully acknowledged.

REFERENCES

Arkes, H. R., and Blumer, C. (1985). The psychology of sunk cost. *Organ. Behav. Hum. Decis. Process.* 35, 124–140. doi: 10.1016/0749-5978(85)90049-4

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *Lme4: Linear Mixed-Effects Models Using Eigen and S4.* Available at: https://cran.r-project.org/web/packages/lme4/index.html

Belschak, F. D., and Den Hartog, D. N. (2009). Consequences of positive and negative feedback: the impact on emotions and extra-role behaviors. *Appl. Psychol. Int. Rev.* 58, 274–303. doi: 10.1111/j.1464-0597.2008.00336.x

Best, D. (2012). *Timeline: Tesco's Attempt to Crack the US. Just-Food*. Available at: https://www.just-food.com/news/tescos-attempt-to-crack-the-us_id121426.aspx [accessed December 5, 2012].

Bonaccio, S., and Dalal, R. S. (2006). Advice taking and decision-making: an integrative literature review, and implications for the organizational sciences. *Organ. Behav. Hum. Decis. Process.* 101, 127–151. doi: 10.1016/j.obhdp.2006.07.001

Brockner, J. (1992). The escalation of commitment to a failing course of action: toward theoretical progress. *Acad. Manage. Rev.* 17, 39–61. doi: 10.2307/258647

Butler, S. (2013). *Tesco Puts US Chain Fresh & Easy in Bankruptcy. The Guardian*. Available at: https://www.theguardian.com/business/2013/oct/01/tesco-us-chain-fresh-easy-bankruptcy [accessed October 1, 2013].

Chakravarty, S., Harrison, G. W., Haruvy, E. E., and Rutström, E. E. (2011). Are you risk averse over other people's money? *South. Econ. J.* 77, 901–913. doi: 10.4284/0038-4038-77.4.901

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Conlon, D. E., and Garland, H. (1993). The role of project completion information in resource allocation decisions. *Acad. Manag. J.* 36, 402–413. doi: 10.2307/256529

Draycott, S., and Dabbs, A. (1998). Cognitive dissonance 1: an overview of the literature and its integration into theory and practice in clinical psychology. *Br. J. Clin. Psychol.* 37, 341–353. doi: 10.1111/j.2044-8260.1998.tb01390.x

Drummond, H. (2014). Escalation of commitment: when to stay the course? *Acad. Manag. Perspect.* 28, 430–446. doi: 10.5465/amp.2013.0039

Elkin, R. A., and Leippe, M. R. (1986). Physiological arousal, dissonance, and attitude change: evidence for a dissonance-arousal link and a "don't remind me" effect. *J. Pers. Soc.* 51, 55–65. doi: 10.1037/0022-3514.51.1.55

Elliot, A. J., and Devine, P. G. (1994). On the motivational nature of cognitive dissonance: dissonance as psychological discomfort. *J. Pers. Soc. Psychol.* 67, 382–394. doi: 10.1037/0022-3514.67.3.382

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Greer, C. R., and Stephens, G. K. (2001). Escalation of commitment: a comparison of differences between Mexican and U.S. decision-makers. *J. Manag.* 27, 51–78. doi: 10.1177/014920630102700104

Han, S., Lerner, J. S., and Keltner, D. (2007). Feelings and consumer decision making: the appraisal-tendency framework. *J. Consum. Psychol.* 17, 158–168. doi: 10.1016/S1057-7408(07)70023-2

Harmon-Jones, E. (2000). Cognitive dissonance and experienced negative affect: evidence that dissonance increases experienced negative affect even in the absence of aversive consequences. *Pers. Soc. Psychol. Bull.* 26, 1490–1501. doi: 10.1177/01461672002612004

Lerner, J. S., and Tiedens, L. Z. (2006). Portait of the angry decision maker: how appraisal tendencies shape anger's influence on cognition. *J. Behav. Decis. Mak.* 19, 115–137. doi: 10.1002/bdm.515

McCain, B. E. (1986). Continuing investment under conditions of failure: a laboratory study of the limits to escalation. *J. Appl. Psychol.* 71, 280–284. doi: 10.1037/0021-9010.71.2.280

Moon, H. (2001a). Looking forward and looking back: integrating completion and sunk-cost effects within an escalation-of-commitment progress decision. *J. Appl. Psychol.* 86, 104–113. doi: 10.1037//0021-9010.86.1.104

Moon, H. (2001b). The two faces of conscientiousness: duty and achievement striving in escalation of commitment dilemmas. *J. Appl. Psychol.* 86, 533–540. doi: 10.1037//0021-9010.86.3.533

O'Boyle, E. H. J., Forsyth, D. R., Banks, G. C., and McDaniel, M. A. (2012). A meta-analysis of the dark triad and work behavior: a social exchange perspective. *J. Appl. Psychol.* 97, 557–579. doi: 10.1037/a0025679.supp

Paulhus, D. L., and Williams, K. M. (2002). The dark triad of personality: narcissism, machiavellianism, and psychopathy. *J. Res. Pers.* 36, 556–563. doi: 10.1016/S0092-6566(02)00505-6

Ronay, R., Oostrom, J. K., Lehmann-Willenbrock, N., and Van Vugt, M. (2017). Pride before the fall: (over)confidence predicts escalation of public commitment. *J. Exp. Soc. Psychol.* 69, 13–22. doi: 10.1016/j.jesp.2016.10.005

Ross, J., and Staw, B. M. (1993). Organizational escalation and exit: lessons from the Shoreham Nuclear Power Plant. *Acad. Manag. J.* 36, 701–732.

Schmidt, J. B., and Calantone, R. J. (2002). Escalation of commitment during new product development. *J. Acad. Mark. Sci.* 30, 103–118. doi: 10.1177/03079459994362

Schoorman, F. D., and Holahan, P. J. (1996). Psychological antecedents of escalation behavior: effects of choice, responsibility, and decision consequences. *J. Appl. Psychol.* 81, 786–794. doi: 10.1037/0021-9010.81.6.786

Schoorman, F. D. (1988). Escalation bias in performance appraisals: an unintended consequence of supervisor participation in hiring decisions. *J. Appl. Psychol.* 73, 58–62. doi: 10.1037/0021-9010.73.1.58

Sleesman, D. J., Conlon, D. E., McNamara, G., and Miles, J. E. (2012). Cleaning up the big muddy: a meta-analytic review of the determinants of escalation of commitment. *Acad. Manag. J.* 55, 541–562. doi: 10.5465/amj.2010.0696

Spielberger, C. D., Krasner, S. S., and Solomon, E. P. (1988). "The experience, expression and control of anger" in *Individual Diferences, Stress, and Health Psychology,* ed. M. P. Janisse (New York, NY: Springer-Verlag New York Inc.), 89–109. doi: 10.1007/978-1-4612-3824-9_5

Staw, B. M. (1976). Knee deep in the big muddy: a study of escalating commitment to a chosen course of action. *Organ. Behav. Hum. Perform.* 16, 27–44. doi: 10.1016/0030-5073(76)90005-2

Staw, B. M., and Fox, F. V. (1977). Escalation: the determinants of commitment to a chosen course of action. *Hum. Relat.* 30, 431–450. doi: 10.1177/001872677703000503

Strough, J., Bruine de Bruin, W., Parker, A. M., Karns, T., Lemaster, P., Pichayayothin, N., et al. (2016). What were they thinking? Reducing sunk-cost bias in a life-span sample. *Psychol. Aging* 31, 724–736. doi: 10.1037/pag0000130

Tata, J. (2002). The influence of managerial accounts on employees ' reactions to negative feedback. *Group Organ. Manag.* 27, 480–503. doi: 10.1177/105960110223858

Tsai, M., and Young, M. J. (2010). Anger, fear, and escalation of commitment. *Cogn. Emot.* 24, 962–973. doi: 10.1080/02699930903050631

Van Overwalle, F., and Jordens, K. (2002). An adaptive connectionist model of cognitive dissonance. *Pers. Soc. Psychol. Rev.* 6, 204–231. doi: 10.1207/S15327957PSPR0603-6

Weiss, H. M., and Cropanzano, R. (1996). "Affective events theory: a theoretical discussion of the structure, causes and consequences of affective experiences at work," in *Research In Organizational Behavior: An Annual Series of Analytical Essays and Critical Reviews,* Vol. 18, eds B. M. Staw and L. L. Cummings (Greenwich, CT: JAI Press Inc.), 1–74.

Werdigier, J. (2013). *Tesco to Pay Dearly to Leave United States. The New York Times*. Available at: https://www.nytimes.com/2013/04/18/business/global/tesco-to-pay-dearly-to-leave-us.html [accessed April 17, 2013].

Wong, K. F., Yik, M., and Kwong, J. Y. (2006). Understanding the emotional aspects of escalation of commitment: the role of negative affect. *J. Appl. Psychol.* 91, 282–297. doi: 10.1037/0021-9010.91.2.282

Wong, K. F., and Kwong, J. Y. (2007). The role of anticipated regret in escalation of commitment. *J. Appl. Psychol.* 92, 545–554. doi: 10.1037/0021-9010.92.2.545

Wong, K. F., and Kwong, J. Y. (2017). Resolving the judgment and decision-making paradox between adaptive learning and escalation of commitment. *Manag. Sci.* 64, 1911–1925. doi: 10.1287/mnsc.2016.2686

Zhang, L., and Baumeister, R. F. (2006). Your money or your self-esteem: threatened egotism promotes costly entrapment in losing endeavors. *Pers. Soc. Psychol. Bull.* 32, 881–893. doi: 10.1177/0146167206287120

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 A

Escalation of Commitment Vignettes

Decision 1

You are the Vice President of Operations for a mid-sized high-tech manufacturing firm. You have 10 million dollars and 3 years to complete a research project that will develop a radar-scrambling device that would render a ship undetectable by conventional radar, in effect, a radar-blank ship. Prior to the beginning of the project, Steve, the project engineer, informs you that he does not think that all 10 million dollars will be needed to successfully complete the project, but he does think that he will need at least 5 million dollars to complete the project.

Between 5 million dollars and 10 million dollars, how much money would you like to invest in the project? $_____

Decision 2

Two years after the project started, Steve retired from the company. Jackie is the new project engineer. You meet with Jackie to get an update on the project. Jackie informs you that Steve used the money you initially invested to purchase inexpensive materials that are of poor quality. As a result, all of the computer components in the plane keep short-circuiting. Jackie says that she is certain she can remedy the mistake, but that she will need an additional 3 million to 6 million dollars in funding. The decision you face now is to either abandon the project or authorize more funding to continue this radar-scrambling project.

Authorize more funding _____

Abandon the project _____

Between 3 million dollars and 6 million dollars, how much money would you like to authorize to continue the radar-scrambling plane? $_____

Decision 3

Three months after you provided the additional funding for Jackie to replace the faulty parts that Steve had purchase, you ask Jackie for an update on the project. You are pleased to learn that the computer components are now working properly. Jackie also informs you that she believes the project will be finished on schedule. However, she informs you that the radar-scrambling device also scrambles other electronic devices, such as the pilot's communication system. She informs you that the problem can be fixed with a new software system for the radar-scrambler, but that she needs an additional 2 million to 5 million funds to purchase the new software system. The decision you face now is to either abandon the project or authorize more funding to continue this radar-scrambling project.

Authorize more funding _____

Abandon the project _____

Between 2 million dollars and 5 million dollars, how much money would you like to authorize to continue the radar-scrambling plane? $_____

Decision 4

After another 3 months have passed, you visit the engineering department to view the radar-scrambling plane. You are pleased to learn that the additional funding you granted solved the problem with the radar-scrambler affecting other devices. Jackie informs you that the plane is ready for a test flight. She asks if you would like to ride aboard the plane during the test flight. You are excited to see how well the plane is working and decide to ride aboard the plane. During the test flight everything works perfectly. None of the radar systems are detecting the plane. 30 min after take-off, the pilot informs you the test is over and he is landing the plane. Once on the ground, you ask the pilot why he landed the plane so shortly after the flight began. He informs you that the additional weight of the radar-scrambling device caused the plane to burn the fuel faster than expected. The pilot suggests that the fuel tanks be upgraded to allow for longer flights but that it would cost an additional 4 million to 7 million dollars in funding. The decision you face now is to either abandon the project or authorize more funding to continue this radar-scrambling project.

Authorize more funding _____

Abandon the project _____

Between 4 million dollars and 7 million dollars, how much money would you like to authorize to continue the radar-scrambling plane? $_____

Decision 5

Three months later, you discover that another firm has already begun marketing a similar product that takes up less space and is much easier to operate than your design. Jackie informs you that the project is 90% complete. She informs you that she is pleased with all of the progress that has been made despite the issues that have arisen along the way. Jackie informs you that although the upgraded fuel tanks allow the plane to fly much further than before, the fuel tanks cost more than expected. She informs you that she will need an additional 1 million to 4 million dollars in funding to pay for the remainder of the

project. The decision you face now is to either abandon the project or authorize more funding to continue this radar-scrambling project.

Authorize more funding _____

Abandon the project _____

Between 1 million dollars and 4 million dollars, how much money would you like to authorize to continue the radar-scrambling plane? $_____

frontiers in Psychology

ORIGINAL RESEARCH
published: 03 August 2018
doi: 10.3389/fpsyg.2018.01203

Does Fear Increase Search Effort in More Numerate People? An Experimental Study Investigating Information Acquisition in a Decision From Experience Task

Jakub Traczyk, Dominik Lenda, Jakub Serek, Kamil Fulawka, Pawel Tomczak, Karol Strizyk, Anna Polec, Piotr Zjawiony and Agata Sobkow*

Wrocław Faculty of Psychology, SWPS University of Social Sciences and Humanities, Wrocław, Poland

OPEN ACCESS

Edited by:
Nathan Dieckmann, Oregon Health & Science University, United States

Reviewed by:
Stephan Dickert, Queen Mary University of London, United Kingdom
Elisabet Tubau, University of Barcelona, Spain

***Correspondence:**
Jakub Traczyk
jtraczyk@swps.edu.pl

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

Received: *18 March 2018*
Accepted: *22 June 2018*
Published: *03 August 2018*

Citation:
Traczyk J, Lenda D, Serek J, Fulawka K, Tomczak P, Strizyk K, Polec A, Zjawiony P and Sobkow A (2018) Does Fear Increase Search Effort in More Numerate People? An Experimental Study Investigating Information Acquisition in a Decision From Experience Task. Front. Psychol. 9:1203. doi: 10.3389/fpsyg.2018.01203

The aim of this study was to investigate the effect of numeracy and the emotion of fear on the decision-making process. While previous research demonstrated that these factors are independently related to search effort, search policy and choice in a decision from experience task, less is known about how their interaction contributes to processing information under uncertainty. We attempted to address this problem and to fill this gap. In the present study, we hypothesized that more numerate people would sample more information about a decision problem and that the effect of fear would depend on the source of this emotion: whether it is integral (i.e., relevant) or incidental (i.e., irrelevant) to a decision problem. Additionally, we tested how these factors predict choices. We addressed these hypotheses in a series of two experiments. In each experiment, we used a sampling paradigm to measure search effort, search policy and choice in nine binary problems included in a decision from experience task. In Experiment 1, before the sampling task we elicited incidental fear by asking participants to recall fearful events from their life. In Experiment 2, integral fear was elicited by asking participants to make choices concerning medical treatment. Decision problems and their payoff distributions were the same in the two experiments and across each condition. In both experiments, we assessed objective statistical numeracy and controlled for a change in the current emotional state. We found that more numerate people sampled more information about a decision problem and switched less frequently between alternatives. Incidental fear marginally predicted search effort. Integral fear led to larger sample sizes, but only among more numerate people. Neither numeracy nor fear were related to the number of choices that maximized expected values. However, across two experiments sample sizes predicted the number of choices that maximized experienced mean returns. The findings suggest that people with higher numeracy may be more sensitive to integral emotions; this may result in more effortful sampling of relevant information leading to choices maximizing experienced returns.

Keywords: numeracy, decision from experience, fear, incidental affect, integral affect, search effort, uncertainty, emotion

INTRODUCTION

In common everyday decision problems (e.g., which financial product to invest in or which drugs to buy to cure a flu) people often do not have explicit information about the full range of possible consequences and their probabilities. Instead, they can acquire sufficient information by actively exploring the structure of a decision problem to select a preferred alternative. The extent to which information about a decision problem is sampled depends on various factors (Wulff et al., 2018). For instance, research has revealed that both numerical abilities (Lejarraga, 2010; Ashby, 2017) and emotions (Frey et al., 2014) are related the exploration process measured by sample sizes in a sampling paradigm of a decision from experience task (Hertwig and Erev, 2009). However, in spite of the fact that emotions can exert a distinct effect on judgment and decision making depending on numerical abilities (Peters et al., 2006b; Peters, 2012), there is scarce research investigating this phenomenon in the context of decisions from experience. The goal of the present study is to fill this gap and to extend our understanding of how numerical abilities and emotions interact in the decision-making process. Namely, we aim to test how objective statistical numeracy and emotion of fear jointly contribute to the exploration of decision problems under uncertainty and whether the amount of acquired information predicts choices. Additionally, we examine whether the source of fear (i.e., integral vs. incidental) may influence the relationship between numeracy and search effort.

Numeracy and Decision Making

Numerous studies have recently documented the advantage of more numerate people in making good decisions (Garcia-Retamero and Cokely, 2017; Cokely et al., 2018). For example, people who are more statistically numerate (i.e., those who better understand the concept of probability and statistical information and are able to use them efficiently; Cokely et al., 2012) are more likely to make normatively superior choices under risk (Pachur and Galesic, 2013) and are less susceptible to some biases (Reyna and Brainerd, 2008; Liberali et al., 2012), which in turn may result in their better actual health (Garcia-Retamero et al., 2015) and accumulation of wealth (Estrada-Mejia et al., 2016).

Among possible cognitive mechanisms underpinning these effects (i.e., selective allocation of attention and judgment calibration; for a detailed discussion see Garcia-Retamero and Cokely, 2017), better decisions in more numerate people are driven at least in part by the fact that such individuals deliberate more on a decision problem (i.e., they spend more time making a choice; Ghazal et al., 2014), acquire more information about potential outcomes as well as their probabilities, and employ more effortful and elaborative processing (Jasper et al., 2017), even if they rarely compute expected values to maximize payoffs (Cokely and Kelley, 2009).

Properties, regarding more extensive and thorough information processing, that characterize people with high numeracy, evidently manifest themselves in making decisions under uncertainty. In laboratory settings, conditions reflecting real-life decision problems with uncertain consequences are often arranged in a decision from experience task (Wulff et al., 2018), in which participants are presented with alternatives (e.g., monetary lotteries). Without any initial knowledge concerning possible outcomes and probabilities, they can freely explore distribution of payoffs to arrive at a decision. To illustrate, in the decision from experience task participants are usually shown with two boxes representing two unknown payoff distributions (e.g., gambles A and B). Participants can freely explore these distributions by uncovering outcomes hidden under each box. For example, in a choice problem in which 3 EUR can be won with the probability of 75% and 5 EUR can be won with the probability of 25% (gamble A) or 5 EUR can be won with the probability of 30% (otherwise nothing; gamble B), participants explore two boxes representing gambles A and B. If a participant draws six samples to sequentially explore gambles A, B, B, B, A, and A, he/she may uncover possible outcomes randomly drawn from the payoff distribution (e.g., 3, 0, 5, 0, 5, and 3; see **Figure 1** for illustration). Search process is terminated, when a participant is ready to make a choice (which gamble A or B he/she prefers). The task can be fully parametrized to investigate a different number of distributions, outcomes, and different levels of probabilities.

To our best knowledge, to date at least two experiments investigated the role of numeracy in information acquisition using the decision from experience task. First, Lejarraga (2010) showed that participants with high numeracy sampled significantly more information across decision problems in comparison to people with low numeracy. Second, Ashby (2017) reported that higher numeracy was related to the larger number of samples drawn and greater consistency in choices across two choice formats (i.e., description vs. experience).

To summarize, these findings draw a picture of a highly numerate individual who deliberates more on a decision problem, exhaustively processes information and enjoys thinking about a decision problem (Lag et al., 2014; Bruine de Bruin et al., 2015). Consequently, in the decision from experience task, such individuals are likely to exhibit more effort in information search resulting in greater engagement in elaborative but also affective encoding of such content in long-term memory representation (Cokely et al., 2018).

Emotions and Decision Making

Throughout the past three decades, increasing attention has been paid to an important role of affect, emotions and feelings in judgment and decision making (Bechara, 2000; Loewenstein and Lerner, 2003; Peters et al., 2006a; Lerner et al., 2015). The focus on emotions resulted in developing various descriptive models positing, among others, that behavior under risk can be driven by feelings (the risk-as-feelings hypothesis; Loewenstein et al., 2001), affect mediates the relationship between risks and benefits (the affect heuristic; Slovic et al., 2007) or that emotions can signal future negative consequences of choices what subsequently helps one to select an advantageous option (the somatic marker hypothesis; Bechara and Damasio, 2005).

Interestingly, it has been documented that the impact of emotions on decision making depends on appraisal tendencies—goal-directed processes that are associated with specific emotions

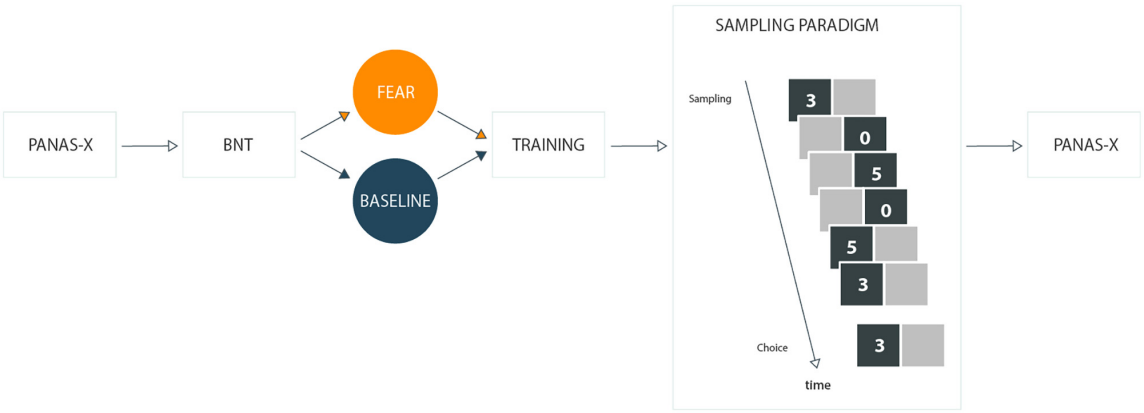


FIGURE 1 | Schematic illustration of the general experimental procedure.

and go beyond their valence only (Lerner and Keltner, 2000). As predicted by the Appraisal-Tendency Framework (Han et al., 2007), different emotions can trigger different cognitive predispositions to assess future events in line with appraisal dimensions that triggered these emotions. To put this in the context of information acquisition, Frey et al. (2014) found the emotion of fear (in comparison to a baseline emotional state) influenced search behavior as measured by sample size and switching frequency between alternatives. At the same time there were no credible effects of sadness and anger. Despite the same negative valence as fear, these emotions probably triggered different appraisal tendencies. The authors argue that fear could have triggered appraisals related to low certainty, high situational control and high anticipated effort which, in turn, evoked a compensatory behavioral response reflected in increased search effort—a response that could have been useful in coping with fear.

Fear is a basic emotion that signals threat in the environment (Öhman and Mineka, 2001) and prepares the survival-related response often operating without conscious experience (LeDoux, 1996). Hence, on the one hand it could serve as a cue indicating it is adaptive to collect more information about a threatening stimulus (e.g., in case of diagnosis of a severe disease, a fearful individual would search for more details regarding drugs and possible treatments). On the other hand, research concerning the impact of fear on attention indicates that task-irrelevant fearful stimuli capture attention (Vuilleumier and Schwartz, 2001; Vuilleumier et al., 2001; Fox et al., 2002), leading to impaired performance in a concurrent task and better encoding of these irrelevant stimuli in memory (Matusz et al., 2015). In case of decision tasks (the probabilistic inference task; Wichary et al., 2016), processing task-irrelevant negative and arousing stimuli (e.g., fearful stimuli) may result in attention-narrowing and focusing on the most important information which was manifested by lower search effort in the decision task (e.g., if I think a stranger on a street intends to harm me, I would not put time and effort exploring possible outcomes of selecting a better credit offer, but rather focus on the imminent danger).

We argue that these potential discrepancies regarding the effects of fear on search behavior (i.e., more vs. less information search) can be attributed to the *source* of this emotion. That is, following Lerner et al. (2015), we introduced the distinction between incidental and integral affect. While the former represents task-irrelevant affect not directly related to a decision problem and which does not guide normatively better decisions (e.g., negative mood caused by bad weather should not influence investment decisions), the latter is directly related to a decision problem and can be relevant in terms of making good choices (e.g., fear of losing money may lead to exploring various offers of savings accounts). In the present study, we expect that the source of emotion of fear (i.e., incidental and integral) would contribute to search effort in different ways, depending on numeracy.

The Role of Emotions in Decision Making Depending on Numeracy

An interesting, however, still underexplored line of research has linked numeracy and decision making to affect (Peters et al., 2006b; Peters, 2012). For instance, Peters et al. (2006b) in a series of experiments tested a theoretical idea according to which more numerate people draw more precise affective meaning from comparison of numbers (i.e., affective precision) that in turn guides their decisions. Nevertheless, the role of affective precision in decision making is not consistent. On the one hand, the authors found that more numerate people were likely to make optimal choices–they preferred a smaller bowl with 10 jelly beans, one of which was a winning jelly bean over a larger bowl with 100 jelly beans containing nine winning jelly beans, probably because they rated feelings of goodness or badness of the former one as more clear. On the other hand, the clarity of feelings was related to suboptimal decisions in a different task. That is, people with high numeracy rated a gamble with a small loss as more attractive than a no-loss gamble and this relationship was partially mediated by affective precision.

We argue that these findings may be explained by focusing on the source of affect. For example, it has been found that people with high numeracy are less prone to incidental affect that is not directly related to a decision problem (Traczyk and Fulawka,

2016). At the same time, these individuals are more sensitive to integral affect that is elicited by a decision problem (Petrova et al., 2014).

To illustrate, in a study by Petrova et al. (2014), participants were informed that they own a camera worth 500 EUR. Depending on the experimental condition, the camera was described in a neutral or affective way, eliciting affect-poor and affect-rich contexts, respectively. Then, participants were asked to declare how much they would pay for insurance against the loss of the camera with a given probability, and to rate emotional reactions to this loss. Based on previous research (Rottenstreich and Hsee, 2001), affect-rich conditions should lead to more distorted probability weighting (reflected by the curvature of the probability weighting function) and, in consequence, lower sensitivity to changes in probability. Indeed, the study indicated that participants were less sensitive to changes in probability in the affect-rich condition and the effect was moderated by numeracy. Importantly, higher numeracy was related to higher variance in reported emotions (i.e., more numerate people reported more differentiated ratings of fear and hope to probabilities irrespective of the experimental condition) that in turn predicted higher sensitivity to probability, suggesting that more numerate participants could have extracted more affective information from probabilities that were relevant and integral to a decision problem.

Traczyk and Fulawka (2016) used a similar procedure to manipulate with irrelevant and incidental affect. In their study, participants were asked to perform two unrelated tasks: the insurance task and the perceptual task. In the insurance task, participants declared how much they would pay for insurance on a coupon worth 500 PLN (Polish Zloty), whereas in the perceptual tasks they were instructed to identify target stimuli in a stream of distractors. Depending on the experimental condition, distractors were either neutral or fearful pictures. The results showed that incidental affect led to a lower sensitivity to changes in probability, but this effect was present only in a group of less numerate participants.

Taken altogether, these results suggest that emotional states (e.g., current fear) of different sources (i.e., integral vs. incidental) can have different impacts on processing probabilities depending on the level of numeracy. It also conforms with the predictions of skilled decision theory (Cokely et al., 2018), according to which statistical numeracy directly increases the precision and calibration of affective responses resulting in affectively charged representative understanding of a decision problem. In other words, integral affect can inform more numerate people about a decision problem and motivate them to deliberate and acquire more information to make good decisions.

Overview and Research Hypotheses

Building on the findings reported above, in the present study we expected that people with higher numeracy would put more effort in the exploration of a decision problem. Furthermore, findings regarding numeracy and affect suggest that search effort in decision problems (i.e., the amount of information acquired) should depend on the source of affect. Specifically, we hypothesized that incidental fear (i.e., a fearful state that is not related to a decision problem directly) would influence the amount of information sampled by people with low (but not high) numeracy because low numerate people are more prone to incidental affect. On the other hand, we predicted that integral fear (i.e., a fearful state elicited by a decision problem) that is meaningful to make a choice would influence search effort only in more numerate participants who are more sensitive to changes in probability and have more differentiated emotional responses to probabilities (i.e., emotional responses correlate more strongly with changes in probability). Moreover, we aimed to explore whether greater search effort would be beneficial for choices. Despite previous research investigated the interaction between numeracy and affect (incidental or integral, separately) using standard lottery tasks (e.g., decisions from description employing static lottery sets), to our best knowledge none of these works addressed the differences between integral and incidental affect and the role of numeracy in dynamic decision from experience tasks. We attempted to fill this gap.

These hypotheses were addressed in a series of two experiments.[1] In each experiment, we employed a decision from experience task (a sampling paradigm) to investigate information search in nine binary decision problems. In Experiment 1, before the sampling task we elicited incidental fear by asking participants to recall fearful events from their life. In Experiment 2, integral fear was elicited by asking participants to make choices concerning medical treatment of a hypothetical disease they were suffering. Decision problems and their payoff distributions were the same in the two experiments and across each condition. In both experiments, we assessed objective statistical numeracy and controlled for a change in the current emotional state.

EXPERIMENT 1

Method

Participants

A total of 118 adult volunteers (69% females, M_{age} = 25.3, SD = 6.3) from the general and student populations participated in an online study for course credit or financial compensation (approximately 10 EUR). Participants received explicit information that the study investigates decision making under uncertainty. Participants provided informed consent prior to the experiment, which was approved by the Ethical Board of the SWPS University of Social Sciences and Humanities.

Materials and Methods

We used the following measures:

The positive and negative affect schedule – expanded form

The Positive and Negative Affect Schedule – Expanded Form (PANAS-X; Watson and Clark, 1999) is a comprehensive mood inventory. PANAS-X is a self-report questionnaire, confirmed

[1]In addition to our main hypotheses regarding search effort, we also measured switching frequency. Following Frey et al. (2014) we use terms *search effort* and *switching frequency* to refer to mean sample size and mean switching frequency which together can describe search/exploratory behavior. However, effects of numeracy and fear on the latter measure were not our main interest in this study.

to reflect the hierarchical structure of affect. It also can be used to assess both short-term and long-term individual differences in affect. The scale consists of a group of words and phrases that describe different feelings and emotions (e.g., "afraid," "frightened," and "cheerful"). We measured current basic negative (e.g., fear) and positive (e.g., joviality) emotional states with items drawn from the Polish adaptation of PANAS-X (Fajkowska and Marszał-Wiśniewska, 2009).

The Berlin Numeracy Test

The Berlin Numeracy Test (BNT; Cokely et al., 2012) is a psychometric instrument that measures risk literacy, statistical numeracy and comprehension of probability. Across numerous studies, the BNT has been shown to be an efficient research tool to measure objective numerical abilities. In the present experiment, we used the test consisting of four items presented to participants in a fixed order.

Decision from experience

In both experiments, we used a decision from experience task to investigate information search in nine binary problems used in previous studies (Frey et al., 2014; see **Table 1**). In each decision problem, two boxes representing two alternatives with unknown distribution of payoffs were displayed. Participants were informed that they could freely explore outcomes and frequencies by drawing random samples from each distribution. Having selected an alternative, an outcome was displayed for 1000 ms. Participants decided by themselves which alternative's distribution they wanted to sample from, when to switch between them and when to terminate exploration. Having finished sampling, participants indicated which alternative they preferred by clicking the "Choose" button below boxes and then again on the chosen box. Prior to the experimental procedure, every participant had to pass through two training trials involving decisions from experience and provide correct responses according to instructions in this task. The purpose of this training session was to familiarize participants with procedure mechanisms (i.e., how to explore two alternatives and to choose one of them), but not to prime them with specific responses in subsequent decision problems. To achieve this goal, instead of numerical values, only geometrical shapes in different colors were selected for payoff distributions. During the training session, tips regarding procedure mechanisms were displayed on a computer screen. In the first training trial, participants were asked to explore two simultaneously presented distributions and select (choose) the one where only circles were present. A slightly different task was provided for the second training trial, where participants were asked to explore two presented distributions and select the one where a pink square was displayed more frequently. After each trial participants received feedback. If they chose a wrong distribution, they had to repeat this training trial until they responded correctly.

Procedure

We randomly assigned participants to one of the two conditions (see **Figure 1**), in which we evoked either incidental fear or baseline emotion (i.e., happiness[2]). At the beginning of the experiment, we measured participants' initial emotional state using the PANAS-X scale. Subsequently, we assessed their numeracy using the BNT. Next, participants had to complete previously described two training trials. Having finished the training session, we introduced a between-subjects manipulation with incidental emotion. That is, based on previous research that has demonstrated that mental imagery systematically evokes emotions on a declarative and physiological level (Holmes and Mathews, 2005; Traczyk et al., 2015; Sobkow et al., 2016), participants were asked to recall and write down life events in which they felt the target emotion (i.e., fearful vs. happy live events). Then, prior to the fourth and seventh decision problem, participants were asked again to recall, but this time just to imagine the previously described life events for 30 s. Finally, to track changes in emotional responses, we again measured participants' current emotional state with the PANAS-X.

Results

Manipulation Check

In order to check whether participants in the two conditions differed in fear ratings, we regressed a post-test score in the fear subscale from the PANAS-X on the experimental condition and the corresponding pretest score. We found that, controlling for pretest scores, mean post-test fear ratings were higher in the fear condition than in the baseline condition, $b = 0.99$, $p = 0.004$. Additional analyses showed that people in the fear condition were less happy in the post-test, $b = -1.39$, $p = 0.003$.

The Effects of Numeracy and Incidental Fear on Search Effort, Search Policy and Choice

We performed four linear regression analyses in which we predicted sample size, switching frequency, the number of choices maximizing expected value (EV) and the number of

[2]The method of inducing incidental affect was inspired by the study by Frey et al. (2014) who manipulated with different emotional states (i.e., fear, sadness, and anger) in comparison to a baseline emotional state of happiness.

TABLE 1 | Nine decision problems based on a study by Frey et al. (2014) that we used in the two experiments.

Decision problem	Payoff distributions		Expected values	
	H	L	H	L
1	4, 0.8	3, 1.0	3.2	3
2	−3, 1.0	−32, 0.1	−3	−3.2
3	−3, 1.0	−4, 0.8	−3	−3.2
4	32, 0.1	3, 1.0	3.2	3
5	32, 0.025	3, 0.25	0.8	0.75
6	3, 1.0	5, 0.55	3	2.75
7	11, 0.35	4, 0.9	3.85	3.6
8	−12, 0.25	−32, 0.1	−3	−3.2
9	−4, 0.25	−3, 0.35	−1	−1.05

H-payoff distribution with the higher expected value (EV); L-payoff distribution with the lower EV. Values represent outcomes and their probabilities (e.g., "4, 0.8" stands for an outcome of +4 with the probability of 80%).

choices maximizing experienced mean returns with the same set of variables: BNT, the experimental condition and their interaction. The BNT was *z*-scored and the experimental conditions were coded as 0.5 (the fear condition) and −0.5 (the baseline condition), so the coefficients that include the effect of the experimental condition (i.e., the main effect and interaction terms) can be directly interpreted as a difference between fear and baseline. Despite skewed distributions of dependent variables, we report results using untransformed data for a more straightforward interpretation of regression coefficients (but our main findings also hold when square root transformation or Poisson regression were applied). The relationships among variables used in Experiment 1 are presented in **Table 2**. There were five people who correctly answered four items of the BNT. Ten participants answered three items; 27 participants answered two items; 30 participants answered one item. Forty three people did not answer any item correctly.

Search effort

We averaged sample sizes across nine decision problems per each participant[3] (M = 12.93, SD = 13.50). That is, we summed up samples per each decision problem for each participant. Then, the arithmetic mean of these samples was calculated for each participant. Next, we regressed mean sample size on numeracy, the experimental condition and their interaction. As predicted, numeracy was positively related to the number of samples drawn, b = 4.13, p < 0.001 (**Figure 2**). Furthermore, we observed a marginally significant effect of incidental fear on the sample size, b = −4.12, p = 0.089.[4] Individuals sampled more in the baseline condition in comparison to the fear condition. We did not find an interaction effect of numeracy and the experimental condition on sample size, b = −1.56, p = 0.520 (see **Table 3** for details).

[3]Three participants were excluded from all analyses in Experiment 1 because they did not draw any sample in any of the problems in the entire experiment.

[4]When we applied a square root transformation to dependent variable, the effect became significant, b = −0.62, p = 0.045. Performing Poisson regression with number of samples drawn from the whole experiment as the criterion variable (which is often employed to count data with a right-skewed distribution) led to similar conclusions, b = −0.34, p < 0.001.

TABLE 2 | The relationships among measures used in Experiment 1.

	1	2	3	4	5	6
1. BNT	–					
2. Mean sample size	0.304***	–				
3. Switching rate	−0.175	−0.427***	–			
4. EV choices	0.013	−0.002	−0.041	–		
5. Experienced mean returns	0.317***	0.34***	0.179	−0.044	–	
6. Change in fear ratings	0.096	0.004	0.013	0.012	−0.101	–

*The change in fear ratings was computed by subtracting pretest fear ratings from post-test fear ratings as measured with PANAS-X. *p < 0.05, **p < 0.01, ***p < 0.001.*

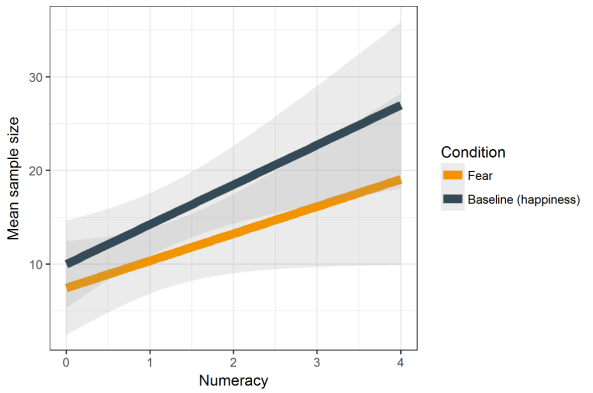


FIGURE 2 | Mean sample size as a function of numeracy and the experimental condition (incidental fear vs. baseline condition – happiness).

Search policy

We analyzed switching behavior following the same procedure as Hills and Hertwig (2010). That is, for each individual and a decision problem we calculated the ratio between the number of actual switches between alternatives and the maximum number of possible switches (i.e., n–1, with n being the total number of drawn samples). Average switching frequency was M = 0.49 (SD = 0.37). Consistent with other findings (Hills and Hertwig, 2012), switching frequency was negatively correlated with sample size, r = −0.43, p < 0.001.

In the regression analysis, we found that BNT was marginally related to switching frequency, b = −0.63, p = 0.071, suggesting that more numerate participants were switching between two alternatives less frequently. The experimental condition and the interaction between numeracy and the experimental condition did not significantly predict the number of switches between alternatives (ps > 0.05).

Expected value maximization

We summed up a total number of choices consistent with the EV maximization principle (M = 3.98, SD = 1.26) to test the effects of numeracy, the experimental condition and their interaction on this measure. None of these predictors significantly explained the number of choices with higher EV (all ps > 0.05).

Experienced mean returns maximization

Hertwig and Pleskac (2008) assumed that people derive their choices from differences in the samples' mean return. The maximization of experienced mean returns predicted more choices in the decision from experience format than expected value in the description format (Wulff et al., 2018). Therefore, we calculated the experienced mean returns summing all the experienced outcomes in the respective decks and dividing them by respective sample sizes. The average number of choices with higher experienced mean returns was M = 5.5 (SD = 2.02). We found that more numerate individuals tended to choose options with higher experienced mean returns, b = 0.66, p < 0.001. Interestingly, we observed the main effect of the experimental condition, b = −0.92, p = 0.010. Participants in the fear condition

made fewer choices consistent with the experienced mean returns maximization than those in the baseline condition. We did not find an interaction effect of numeracy and group (see **Table 3**).

We further explored these effects by introducing mean sample size to the model (**Table 4**). The effects of BNT and fear did hold. Additionally, we found that the number of choices maximizing experienced mean returns was predicted by mean sample size, $b = 0.51$, $p = 0.008$.

Summary

To summarize, results of Experiment 1 supported our main hypothesis, according to which more numerate participants will exhibit more search effort manifested in larger sample sizes. However, opposite to our predictions, incidental fear did not influence search effort in case of people with low numeracy. That is, despite a weak main effect of incidental fear on search effort, numeracy did not moderate this relationship. Moreover, we observed that numeracy, incidental fear and search effort (as measured by mean sample size) predicted the number of choices maximizing experienced mean return. It suggests that more numerate people sampled more information about a decision problem and, in turn, chose more advantageous alternatives based on experienced outcomes.

In Experiment 2, we introduced a procedure modification to test these relationships for a different source of fear: integral fear.

EXPERIMENT 2

Method

Participants

Ninety adult volunteers (60% females, $M_{age} = 26.4$, $SD = 6.7$) took part in Experiment 2. Participants were recruited from the same population as in Experiment 1; they were incentivized in the same way and received the same initial information about the study. The experiment was approved by the Ethical Board of the SWPS University of Social Sciences and Humanities.

Materials and Methods

We used the same materials and method as in Experiment 1.

Procedure

To elicit integral fear, we randomly assigned participants to one of the two between-subjects conditions: medical decisions or financial decisions. In general, we expected the former condition to be relatively more frightening than the latter one because of affect-rich medical outcomes (Pachur et al., 2014; Suter et al., 2015). In Experiment 2, we used almost the same procedure as in Experiment 1. At the outset, we measured initial participants' current emotional state with the PANAS-X. Then we measured numeracy with the BNT, followed by the previously described training trials and decision problems.

The only difference in Experiment 2 was a method of inducing a desirable emotional state. Instead of incidental fear, we used integral fear manipulation. In the medical condition,

TABLE 3 | Results of regression analyses predicting sample size, switching frequency and choices maximizing EV in Experiment 1.

Predictor	Sample size			Switching frequency			EV choices			Experienced mean returns		
	b	95% CI	*p*	*b*	95% CI	*p*	*b*	95% CI	*p*	*b*	95% CI	*p*
Intercept	12.90**	(10.52, 15.27)	<0.001	0.49**	(0.42, 0.56)	<0.001	3.99**	(3.76, 4.22)	<0.001	5.49	(5.1, 5.83)	<0.001
Numeracy	4.13**	(1.74, 6.51)	<0.001	−0.06†	(−0.13, 0.01)	0.071	0.01	(−0.23, 0.24)	0.954	0.66***	(0.31, 1.01)	<0.001
Group (Fear)	−4.12†	(−8.86, 0.64)	0.089	−0.04	(−0.18, 0.10)	0.632	0.28	(−0.18, 0.74)	0.238	−0.92*	(−1.6,−0.22)	0.010
Numeracy*Group	−1.56	(−6.32, 3.22)	0.520	0.1	(−0.04, 0.24)	0.134	−0.32	(−0.78, 0.16)	0.186	0.46	(−0.24, 1.16)	0.190
		$R^2 = 0.119$**			$R^2 = 0.052$			$R^2 = 0.028$			$R^2 = 0.165$***	

Group was coded as 0.5 (Fear) and −0.5 (Baseline). †*p < 0.1, *p < 0.05, **p < 0.01, ***p < 0.001.*

TABLE 4 | Results of regression analyses predicting experienced mean returns in Experiment 1 and Experiment 2.

Predictor	Experiment 1			Experiment 2		
	b	95% CI	*p*	*b*	95% CI	*p*
Intercept	5.53	(5.16, 5.90)	<0.001	5.71	(5.19, 6.23)	<0.001
Numeracy	0.50*	(0.11, 0.89)	0.012	0.06	(−0.52, 0.63)	0.841
Group (Fear)	−0.74†	(−1.48, 0.00)	0.051	−0.38	(−1.40, 0.66)	0.470
Sample size	0.51**	(0.13, 0.89)	0.008	1.22**	(0.40, 2.04)	0.004
Numeracy*Group	0.42	(−0.36, 1.18)	0.296	0.42	(−0.72, 1.56)	0.470
Numeracy*Sample size	−0.11	(−0.54, 0.33)	0.628	−0.25	(−1.22, 0.71)	0.603
Group*Sample size	0.16	(−0.58, 0.92)	0.661	−0.84	(−2.48, 0.78)	0.307
Numeracy*Group*Sample size	−0.10	(−0.96, 0.78)	0.827	−0.68	(−2.62, 1.24)	0.480
		$R^2 = 0.220$***			$R^2 = 0.134$***	

Group was coded as 0.5 (Fear) and −0.5 (Baseline). †*p < 0.1, *p < 0.05, **p < 0.01, ***p < 0.001.*

participants were asked to imagine they were diagnosed with a dangerous, lethal disease. During the experimental procedure, they had to choose one of two drugs (choice alternatives), which could accelerate or slow down the development of the disease, based on an underlying distribution of payoffs. In other words, a distribution of payoffs in the decision task indicated how many years a treatment with a particular drug could prolong or shorten life given that a disease was diagnosed (e.g., an outcome of "3" informed participants that a patient randomly drawn from a population of patients who underwent a treatment with a particular drug lived 3 years longer than expected; see **Supplementary Table S1** for the exact instructions). In the financial condition, participants were asked to imagine they are CEOs of a new company and their task is to choose between two financial products, which could be beneficial or disadvantageous investments from the perspective of the company. As in the medical condition, potential outcomes of financial products were hidden under two boxes representing payoff distributions of the two alternative investments. In both conditions, participants faced nine decision problems with the same payoff distributions as in Experiment 1. At the end, we once again measured their current emotional state with the PANAS-X.

Results

All analyses reported below followed the approach we employed in Experiment 1. One participant did not complete the whole procedure and was excluded from further analyses.[5] The relationships among variables used in Experiment 2 are presented in **Table 5**. There were nine people who correctly answered four items of the BNT. Nine participants answered three items; 22 participants answered two questions; 23 participants answered one item. Twenty five people did not answer any item correctly.

[5]The overall pattern of results did not change when we included data from this participant.

TABLE 5 | The relationships among measures used in Experiment 2.

	1	2	3	4	5	6
1. BNT	–					
2. Mean sample size	0.328**	–				
3. Switching rate	−0.286**	−0.44***	–			
4. EV choices	0.099	0.095	−0.029	–		
5. Experienced mean returns	0.08	0.163	0.059	0.009	–	
6. Change in fear ratings	0.142	−0.067	−0.103	0.075	−0.041	–

*Change in fear ratings was computed by subtracting pretest fear ratings from post-test fear ratings as measured with PANAS-X. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.*

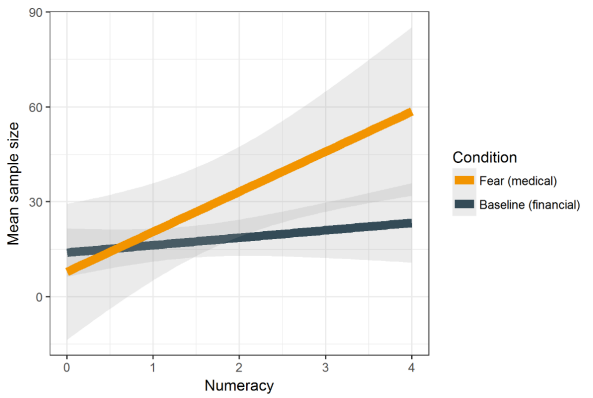


FIGURE 3 | Mean sample size as a function of numeracy and the experimental condition (integral fear – medical problems vs. baseline condition – financial problems).

Manipulation Check

To check the effectiveness of the experimental manipulation, we regressed a post-test score in the fear scale from PANAS-X on the experimental conditions (i.e., medical vs. financial outcomes) and the corresponding pretest score. Controlling for pretest scores, mean post-test fear ratings were higher in the medical condition than in the financial condition, $b = 1.86$, $p = 0.019$, which supports our prediction that medical outcomes would evoke more fearful responses in comparison to financial outcomes. We did not observe significant differences in the level of happiness between the medical condition and the financial condition, $b = -0.32$, $p = 0.652$.

The Effects of Numeracy and Integral Fear on Search Effort, Search Policy and Choice

Search effort

We averaged sample sizes across nine decision problems from each participant ($M = 22.14$, $SD = 32.36$). Replicating our previous findings, we found that the BNT was positively related to sample size, $b = 9.67$, $p = 0.003$. There was no main effect of the experimental condition, $b = 9.16$, $p = 0.165$. However, we found a significant interaction of these predictors, $b = 13.28$, $p = 0.044$ (**Figure 3**). As we expected, a simple effects analysis performed for each level of numeracy confirmed that participants with higher numeracy (those who gave correct responses in three or four items in the BNT) explored payoff distributions to a greater extent in the integral fear medical condition in comparison to the baseline financial condition ($p = 0.043$ and $p = 0.030$ for people correctly answering three and four items in the BNT, respectively).

Additionally, we tested whether sample sizes differed between Study 1 and Study 2. We found that in Study 2 participants drew significantly more samples (22 samples) than in Study 1 (13 samples), $t(110.23) = -2.5$, $p = 0.013$. This suggests that integral affect (Study 2), in comparison to incidental affect (Study 1), might have increased search effort.

TABLE 6 | Results of regression analyses predicting sample size, switching frequency, choices maximizing EV and experienced mean returns in Experiment 2.

	Sample size			Switching frequency			EV choices			Experienced mean returns		
Predictor	***b***	**95% CI**	***p***	***b***	**95% CI**	***p***	***b***	**95% CI**	***p***	***b***	**95% CI**	***p***
Intercept	22.01**	(15.51, 28.52)	<0.001	0.40**	(0.32, 0.48)	<0.001	4.46**	(4.11, 4.81)	<0.001	5.47	(4.98, 5.97)	<0.001
Numeracy	9.67**	(3.20, 16.13)	0.003	−0.11*	(−0.19,−0.03)	0.011	0.17	(−0.17, 0.52)	0.325	0.2	(−0.3, 0.69)	0.430
Group (Fear)	9.16	(−3.84, 22.18)	0.165	−0.08	(−0.24, 0.08)	0.363	−0.28	(−0.98, 0.40)	0.417	−0.38	(−1.38, 0.62)	0.453
Numeracy*Group	13.28*	(0.36, 26.22)	0.044	0.00	(−0.16, 0.18)	0.915	0.12	(−0.58, 0.80)	0.742	0.44	(−0.54, 1.44)	0.373
		R^2 = 0.169**			R^2 = 0.091			R^2 = 0.019			R^2 = 0.022	

*Group was coded as 0.5 (Fear) and −0.5 (Baseline). *p < 0.05, **p < 0.01, ***p < 0.001.*

Search policy

Average switching frequency was $M = 0.41$ ($SD = 0.39$). Results of regression indicated that participants with higher numeracy alternated back and forth between lotteries less frequently, $b = -0.11$, $p = 0.011$. Nonetheless, the experimental condition and its interaction with BNT were not significant predictors of switching frequency ($ps > 0.5$). Moreover, we corroborated results from Experiment 1, showing that a correlation between total sample size and switching frequency was negative, $r = -0.44$, $p < 0.001$.

Expected value maximization

On average, participants selected $M = 4.49$ ($SD = 1.59$) alternatives with higher EV. As with Experiment 1, no predictors were significantly related to the number of decisions consistent with EV maximization principle (all $ps > 0.05$).

Experienced mean returns maximization

We calculated experienced mean returns following the same rationale as in the previous experiment. Average number of choices with higher experienced mean returns was $M = 5.53$ ($SD = 2.28$). We did not replicate findings from Experiment 1. None of the predictors substantially explained variance of the experienced mean returns (see **Table 6**).

When mean sample size was introduced to the model, we again found that choices maximizing experienced mean returns significantly were predicted by mean sample size, $b = 1.22$, $p = 0.004$ (**Table 4**).

Summary

We found that numeracy was the strongest predictor of search effort. Interestingly, although integral fear evoked by medical problems did not influence search effort directly, this relationship was moderated by numeracy. In particular, participants with higher numeracy tended to draw more samples in medical decision problems that elicited a greater integral fear. However, none of the predictors were related to choices maximizing EV. Nevertheless, sample size, but not numeracy or the experimental condition, predicted the number of choices maximizing experienced mean returns. Therefore, the relationship between numeracy and choice was not as straightforward as in Experiment 1. The pattern of results suggests that in the case of the integral fear manipulation, numeracy may influence choices indirectly by increasing search effort.

GENERAL DISCUSSION

The goal of the present study was to investigate the role of numeracy and emotions in decision-making process. Specifically, we used a decision from experience task to measure search effort as a function of statistical numeracy and fear. In two experiments, we found that numeracy is a robust predictor of search effort. In particular, more numerate participants tended to acquire more information about outcomes and their probabilities. Interestingly, numeracy moderated the relationship between fear and search effort but only when the source of this emotion was integral to a decision problem. That is, when fear was produced by outcomes and probabilities of the decision problem rather than being elicited by a concurrent and unrelated task, more numerate people draw more samples in comparison to people with low numeracy. These findings imply that people with high numeracy may use relevant affect as an additional cue when processing information about a decision problem. Additionally, we demonstrated that numeracy and fear did not increase the number of choices that maximized EV, but they predicted choices maximizing the experienced mean returns. Nevertheless, the nature of this relationship is not straightforward. That is, numeracy was directly related to choices in Experiment 1, while in Experiment 2, numeracy may have operated indirectly through increased search effort.

The Adaptive Role of Fear and Numeracy in Information Acquisition

Fear, as one of the basic emotions present among humans and other primates (Ekman, 1992), informs an organism about dangers and prepares responses to a potential environmental threat (Öhman and Mineka, 2001). In this sense fear plays an adaptive and survival-related function, as it focuses attention on a threatening stimulus and motivates an organism to cope with the threat or avoid it (Adolphs, 2013). A threatening stimulus can trigger a pattern of conditioned and unconditioned behavioral responses to danger or activate "survival circuits" responsible for producing an adaptive defense response to such stimuli (LeDoux, 2014). However, the automatic survival-related response to fear may become maladaptive if it holds attentional resources and prevents an individual from processing a more important concurrent task not directly related to a threatening stimulus.

In case of decision-making research, reported effects of negative affect (e.g., fear) on process-tracing measures are mixed. For example, emotional stress (evoked by task-irrelevant high-arousal pictures of negative valence) reduced information search and promoted simplified processing (Wichary et al., 2016). Similarly, choices in the affect-rich medical domain (i.e., the choice between two medications with negative side-effects of different severity) were associated with lower sample sizes in comparison to the monetary domain (Lejarraga et al., 2016). On the other hand, fearful participants put more effort in the exploration of payoff distributions (Frey et al., 2014) in comparison to the baseline happy condition. This supports the notion that the negative affect is associated with more in-depth and elaborate information processing (for a theoretical account see Forgas, 1995; Schwarz, 2001).

Conclusions from our study suggest there are at least two important factors that offer an opportunity to explain inconsistent findings regarding the role of fear (and more generally emotions) in searching information about a decision problem. These factors are the source of fear and numeracy.

The adaptive and informational advantage of fear is contingent on whether the fear is relevant or irrelevant to a decision problem. Such distinctions have already been noticed in communication and persuasion studies showing that a negative emotional state (e.g., fear) increased motivation to elaborate when the threat was relevant and integral to a stimulus, but decreased motivation to elaborate on information that is incidental and irrelevant (Baron et al., 1994). The distinction between incidental and integral affect/emotions has also been raised in judgment and decision-making research (e.g., Lerner et al., 2015; Västfjäll et al., 2016). In line with these theoretical considerations, our study demonstrated that a state of fear elicited incidentally by an unrelated task led to lower search effort. This effect could have been driven by fearful mental images that captured and absorbed people's attention, drawing it away from the exploration of choice alternatives. Moreover, fear serves as a cue in goal queuing (Simon, 1967): It may interrupt the current program and give high priority to real-time needs. In this case, people could have been more motivated to cope with fearful mental images instead of performing experimental tasks.

On the contrary, the state of fear elicited by the properties of a decision task could have focused people's attention and motivation on the decision problem, increasing search effort and decreasing the switching rate. Interestingly, integral fear had an impact on search effort only among more numerate participants who are more sensitive to number-related affective reactions (Peters, 2012; Petrova et al., 2014; Peters and Bjalkebring, 2015) and are able to derive a richer gist of information (Reyna, 2004; Reyna et al., 2009). Consequently, integral fear can be adaptive in such problems because it gives more priority to the experimental task and influences search effort as well as a more extensive exploration of important features of a decision problem. Since we did not test attentional engagement in these tasks, future studies using eye tracking or psychophysiological methods could address this hypothesis directly.

The Prediction of Choices

It has recently been theorized (Cokely et al., 2018), as well as empirically confirmed (Jasper et al., 2013; Traczyk et al., 2018) that people with high numeracy exhibit more adaptive behavior under risk and uncertainty. At the level of information processing, we can conclude that more numerate people in our study were adaptive: They put in more effort when fear was integral to the problem, yet were not influenced by incidental fear. Under such conditions using integral fear as information is adaptive, as it allows one to acquire more information about the important problem before making a final decision. At the level of choice, the question of adaptive behavior is more complicated because of the variety of criteria defining good choices (Hogarth, 2015).

In our study, higher numeracy did not predict choices maximizing EV. This is not surprising in light of other research that showed the lack of such a relationship (Ashby, 2017) or a moderating role of other factors in the relationship between numeracy and choices maximizing EV (Traczyk et al., 2018). Moreover, in contrast to previous research (Frey et al., 2014), we did not observe the effect of fear on maximizing EV. These results can in part be explained by the fact that people who took part in our experiments were not paid contingent on their actual choices. Another explanation is that more numerate people followed different criteria of good choices instead of normative standards (e.g., maximizing expected value or expected utility). Accordingly, they could have drawn "enough" samples to make a satisfactory choice that maximized returns based on experienced outcomes. Additionally, it has been demonstrated that more numerate people often do not compute EV of a gamble but rather employ elaborative heuristic processing (Cokely and Kelley, 2009). That is, people with high numeracy consider more aspects of a choice problem, they recode probabilities, focus on maximum and minimum differences between outcomes or take their risk preference into account. This may result in longer deliberation about problem leading to superior choices.

Across two experiments, numeracy predicted search effort. This was positively related to more choices maximizing the experienced mean returns. It suggests that search effort may be a key factor explaining good choices. Nevertheless, the relationship between these measures is not straightforward. In Experiment 1 numeracy predicted choices maximizing experienced mean returns directly irrespective of incidental fear, while in Experiment 2 participants with high numeracy were likely to sample more information, which successively influenced choices.

An interesting question that emerges from our findings addresses the role of numeracy in search/exploratory behavior in general. In the manuscript, we reported that higher sampling was related to lower switching rate (people who generally sampled more also switched less between alternatives). In case of numeracy, Ashby (2017) showed that higher numeracy was related to lower switching rate. Our additional analyses corroborated this result. It may imply that more numerate people extensively explored only one of the two options. However, it is more plausible that people (particularly more numerate individuals) used rather a comprehensive strategy of sampling

than a piecewise strategy. The piecewise search oscillates between options, each time drawing the smallest possible sample. On the other hand, a decision-maker who applies the comprehensive policy samples extensively from one option and then samples extensively from the other option. Therefore, highly numerate people could less frequently alternate back and forth between options, but, at the same time, it does not necessarily mean that they tended to sample more from one option while ignoring the other, because this alternating could be equal across all options.

We believe these findings may contribute to better understanding of the role of numeracy and fear in decision making, but also may have some practical applications. Further studies may precisely tackle this issue. For example, it seems appealing to investigate methods of increasing search effort in less numerate people by directing their attention to integral, fear-related aspects of a decision problem. Such interventions could support less numerate people in sampling more information resulting in better choices. The emotion-based intervention might be a complementary to other aids (e.g., visual aids) designed to improve risk literacy.

Limitations and Future Research

Although we tried to minimize differences in experimental procedures between Study 1 and Study 2 in order to compare the effects of incidental and integral affect, there are still some concerns regarding comparability of our tasks. That is, keeping the same payoff distributions and choice problems, the procedures differed in instructions provided to participants. In Study 1 participants were instructed to imagine events from their life while in Study 2, we only manipulated the content of instructions (e.g., financial vs. medical scenarios). As a result, we found that integral but not incidental affect influenced people with high numeracy who searched for more information. Additionally, we demonstrated that in general, integral affect was related to more search effort in comparison to incidental affect. This suggests that integral affect is likely to increase search effort. However, an alternative explanation is also plausible. The differences in procedures might have influenced motivation or engagement that led to more extensive search. Moreover, one could argue that the experimental manipulation in Study 2 was not a direct manipulation of integral fear and it could have also influenced (or activated) other constructs that may be potentially related to motivation (e.g., mortality salience in the medical condition or money priming in the financial condition, Zaleskiewicz et al., 2013). This issue could be addressed by using other methods to compare medical and financial gambles (Lejarraga et al., 2016) and to control for potentially confounding variables.

Another interesting line of future research is to investigate whether more numerate people are more sensitive to integral emotions or rather they are able to experience integral emotions more accurately because of their previous personal experience. For instance, if a person had more personal experience with financial decisions that resulted in losses than with medical decisions, he/she may be more sensitive to such affective influences or experience them with more intensity. Furthermore, numeracy may moderate this relationship.

Finally, because numeracy has not been experimentally manipulated in our study (for an example of a study with manipulation intended to improve numeracy see Peters et al., 2017), drawing causal inferences about the influence of numeracy on decision making seems problematic. Numeracy is correlated with many measures such as intelligence (Lag et al., 2014), need for cognition (Bruine de Bruin et al., 2015), cognitive reflection (Weller et al., 2013) and education (Ghazal et al., 2014), so it cannot be excluded that one of these factors was responsible for the effects obtained in the current study. Nevertheless, previous research has demonstrated that the effects of numeracy hold even when controlling for the above mentioned variables, suggesting that numeracy is a robust unique predictor of superior decision making (Cokely et al., 2018).

Conclusion

To summarize, our study demonstrated that people with high numeracy acquire more information about a decision problem. Importantly, more numerate people seem to use task-relevant affective information as a cue signaling the importance of a decision problem. This in turn motivates them to put more effort in the exploration of outcomes and their probabilities. In consequence of greater search effort, people with high numeracy are able to maximize experienced mean return. Altogether, decisions made by highly numerate people may be guided not only by objective properties of choice problems (e.g., outcomes), but also by adaptive affective responses to these problems.

DATA AVAILABILITY STATEMENT

All analyzed datasets for this study are included in the manuscript and the **Supplementary Files**.

FUNDING

The research in this article were supported by the National Science Centre, Poland under grant 2015/17/D/HS6/00703 and the Foundation for Polish Science (FNP) under grant START (111.2016).

AUTHOR CONTRIBUTIONS

JT developed the study concept and acquired funding. JT, AS, DL, JS, PT, KF, AP, KS, and PZ contributed equally to the study design. DL, JS, and JT programmed the procedure. DL and JT performed the data analysis. JT, AS, DL, JS, PT, KF, AP, KS, and PZ conducted the study, wrote and reviewed the article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01203/full#supplementary-material

DATA SHEET S1 | Data set - Experiment 1.

DATA SHEET S2 | Data set - Experiment 2.

DATA SHEET S3 | Description of data sets and variables.

TABLE S1 | Instructions presented to participants in two experiments.

REFERENCES

Adolphs, R. (2013). The biology of fear. *Curr. Biol.* 23, R79–R93. doi: 10.1016/j.cub.2012.11.055

Ashby, N. J. S. (2017). Numeracy predicts preference consistency: deliberative search heuristics increase choice consistency for choices from description and experience. *Judgm. Decis. Mak.* 12, 128–139.

Baron, R., Logan, H., Lilly, J., Inman, M., and Brennan, M. (1994). Negative emotion and message processing. *J. Exp. Soc. Psychol.* 30, 181–201. doi: 10.1006/jesp.1994.1009

Bechara, A. (2000). Emotion, decision making and the orbitofrontal cortex. *Cereb. Cortex* 10, 295–307. doi: 10.1093/cercor/10.3.295

Bechara, A., and Damasio, A. R. (2005). The somatic marker hypothesis: a neural theory of economic decision. *Games Econ. Behav.* 52, 336–372. doi: 10.1016/j.geb.2004.06.010

Bruine de Bruin, W., McNair, S. J., Taylor, A. L., Summers, B., and Strough, J. (2015). "Thinking about numbers is not my idea of fun": need for cognition mediates age differences in numeracy performance. *Med. Decis. Mak.* 35, 22–26. doi: 10.1177/0272989X14542485

Cokely, E. T., Feltz, A., Ghazal, S., Allan, J. N., Petrova, D., and Garcia-Retamero, R. (2018). "Decision making skill: From intelligence to numeracy and expertise," in *The Cambridge Handbook of Expertise and Expert Performance*, eds K. A. Ericsson, R. R. Hoffman, A. Kozbelt, and A. M. Williams (New York, NY: Cambridge University Press), 476–505.

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. doi: 10.1177/0272989X16655334

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.

Ekman, P. (1992). An argument for basic emotions. *Cogn. Emot.* 6, 169–200. doi: 10.1080/02699939208411068

Estrada-Mejia, C., de Vries, M., and Zeelenberg, M. (2016). Numeracy and wealth. *J. Econ. Psychol.* 54, 53–63. doi: 10.1016/j.joep.2016.02.011

Fajkowska, M., and Marszał-Wiśniewska, M. (2009). Właściwości psychometryczne Skali Pozytywnego i Negatywnego Afektu-Wersja Rozszerzona (PANAS-X). Wstępne wyniki badań w Polskiej próbie. *Przedgląd Psychol.* 52, 355–387.

Forgas, J. P. (1995). Mood and judgment: the affect infusion model (AIM). *Psychol. Bull.* 117, 39–66. doi: 10.1037/0033-2909.117.1.39

Fox, E., Russo, R., and Dutton, K. (2002). Attentional bias for threat: evidence for delayed disengagement from emotional faces. *Cogn. Emot.* 16, 355–379. doi: 10.1080/02699930143000527

Frey, R., Hertwig, R., and Rieskamp, J. (2014). Fear shapes information acquisition in decisions from experience. *Cognition* 132, 90–99. doi: 10.1016/j.cognition.2014.03.009

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

Garcia-Retamero, R., and Cokely, E. T. (2017). Designing visual aids that promote risk literacy: a systematic review of health research and evidence-based design heuristics. *Hum. Factors* 59, 582–627. doi: 10.1177/0018720817690634

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.

Han, S., Lerner, J. S., and Keltner, D. (2007). Feelings and consumer decision making: the appraisal-tendency framework. *J. Consum. Psychol.* 17, 158–168. doi: 10.1016/S1057-7408(07)70023-2

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 Pleskac, T. J. (2008). "The game of life: How small samples render choice simpler," in *The Probabilistic Mind: Prospects for Rational Models of Cognition*, eds N. Charter and M. Oaksford (Oxford: Oxford University Press), 209–236. doi: 10.1093/acprof:oso/9780199216093.003.0010

Hills, T. T., and Hertwig, R. (2010). Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychol. Sci.* 21, 1787–1792. doi: 10.1177/0956797610387443

Hills, T. T., and Hertwig, R. (2012). Two distinct exploratory behaviors in decisions from experience: comment on Gonzalez and Dutt (2011). *Psychol. Rev.* 119, 888–892. doi: 10.1037/a0028004

Hogarth, R. M. (2015). "What's a 'Good' Decision? Issues in assessing procedural and ecological quality," in *The Wiley Blackwell Handbook of Judgment and Decision Making*, eds K. Gideon and G. Wu (Chichester: John Wiley & Sons, Ltd.), 952–972. doi: 10.1002/9781118468333.ch34

Holmes, E. A., and Mathews, A. (2005). Mental imagery and emotion: a special relationship? *Emotion* 5, 489–497. doi: 10.1037/1528-3542.5.4.489

Jasper, J. D., Bhattacharya, C., and Corser, R. (2017). Numeracy predicts more effortful and elaborative search strategies in a complex risky choice context: a process-tracing approach. *J. Behav. Decis. Mak.* 30, 224–235. doi: 10.1002/bdm.1934

Jasper, J. D., Bhattacharya, C., Levin, I. P., Jones, L., and Bossard, E. (2013). Numeracy as a predictor of adaptive risky decision making. *J. Behav. Decis. Mak.* 26, 164–173. doi: 10.1002/bdm.1748

Lag, 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

LeDoux, J. E. (1996). *The Emotional Brain: The Mysterious. Underpinning of Emotional Life*. New York, NY: Simon & Schuster.

LeDoux, J. E. (2014). Coming to terms with fear. *Proc. Natl. Acad. Sci. U.S.A.* 111, 2871–2878. doi: 10.1073/pnas.1400335111

Lejarraga, T. (2010). When experience is better than description: time delays and complexity. *J. Behav. Decis. Mak.* 23, 100–116. doi: 10.1002/bdm.666

Lejarraga, T., Pachur, T., Frey, R., and Hertwig, R. (2016). Decisions from experience: from monetary to medical gambles. *J. Behav. Decis. Mak.* 29, 67–77. doi: 10.1002/bdm.1877

Lerner, J. S., and Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cogn. Emot.* 14, 473–493. doi: 10.1080/026999300402763

Lerner, J. S., Li, Y., Valdesolo, P., and Kassam, K. S. (2015). Emotion and decision making. *Annu. Rev. Psychol.* 66, 799–823. doi: 10.1146/annurev-psych-010213-115043

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., and Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *J. Behav. Decis. Mak.* 25, 361–381. doi: 10.1002/bdm.752

Loewenstein, G., and Lerner, J. S. (2003). "The role of affect in decision making," in *Handbook of Affective Sciences*, eds R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (New York, NY: Oxford University Press), 619–642.

Loewenstein, G., Weber, E. U., Hsee, C. K., and Welch, N. (2001). Risk as feelings. *Psychol. Bull.* 127, 267–286. doi: 10.1037//0033-2909.127.2.267

Matusz, P. J., Traczyk, J., Sobkow, A., and Strelau, J. (2015). Individual differences in emotional reactivity moderate the strength of the relationship between attentional and implicit-memory biases towards threat-related stimuli. *J. Cogn. Psychol.* 27, 715–724. doi: 10.1080/20445911.2015.1027210

Öhman, A., and Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychol. Rev.* 108, 483–522. doi: 10.1037//0033-295X.108.3.483

Pachur, T., and Galesic, M. (2013). Strategy selection in risky choice: the impact of numeracy, affect, and cross-cultural differences. *J. Behav. Decis. Mak.* 26, 260–271. doi: 10.1002/bdm.1757

Pachur, T., Hertwig, R., and Wolkewitz, R. (2014). The affect gap in risky choice: affect-rich outcomes attenuate attention to probability information. *Decision* 1, 64–78. doi: 10.1037/dec0000006

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., and Bjalkebring, P. (2015). multiple numeric competencies: when a number is not just a number. *J. Pers. Soc. Psychol.* 108, 802–822. doi: 10.1037/pspp0000019

Peters, E., Shoots-reinhard, B., Tompkins, M. K., Schley, D., Meilleur, L., Sinayev, A., et al. (2017). Improving numeracy through values affirmation enhances decision and STEM outcomes. *PLoS One* 12:e0180674. doi: 10.1371/journal.pone.0180674

Peters, E., Västfjäll, D., Gärling, T., and Slovic, P. (2006a). Affect and decision making: a "hot" topic. *J. Behav. Decis. Mak.* 19, 79–85. doi: 10.1002/bdm.528

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

Petrova, D., van der 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

Reyna, V. F. (2004). How people make decisions that involve risk. *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

Rottenstreich, Y., and Hsee, C. K. (2001). Money, kisses, and electric shocks: on the affective psychology of risk. *Psychol. Sci.* 12, 185–190. doi: 10.1111/1467-9280.00334

Schwarz, N. (2001). "Feelings as information: implications for affective influences on information processing," in *Theories of Mood and Cognition: A User's Handbook*, eds L. L. Martin and G. L. Clore (Mahwah, NJ: Erlbaum), 159–176.

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychol. Rev.* 74, 29–39. doi: 10.1037/h0024127

Slovic, P., Finucane, M. L., Peters, E., and MacGregor, D. G. (2007). The affect heuristic. *Eur. J. Oper. Res.* 177, 1333–1352. doi: 10.1016/j.ejor.2005.04.006

Sobkow, A., Traczyk, J., and Zaleskiewicz, T. (2016). The affective bases of risk perception: negative feelings and stress mediate the relationship between mental imagery and risk perception. *Front. Psychol.* 7:932. doi: 10.3389/fpsyg.2016.00932

Suter, R. S., Pachur, T., and Hertwig, R. (2015). How affect shapes risky choice: distorted probability weighting versus probability neglect. *J. Behav. Decis. Mak.* 29, 437–449. doi: 10.1002/bdm.1888

Traczyk, J., and Fulawka, K. (2016). Numeracy moderates the influence of task-irrelevant affect on probability weighting. *Cognition* 151, 37–41. doi: 10.1016/j.cognition.2016.03.002

Traczyk, J., Sobkow, A., Fulawka, K., Kus, J., Petrova, D., and Garcia-Retamero, R. (2018). Numerate decision makers don't use more effortful strategies unless it pays: a process tracing investigation of skilled and adaptive strategy selection in risky decision making. *Judgm. Decis. Mak.* 13, 372–381. Available at: http://www.sjdm.org/journal/17/17208/jdm17208.pdf

Traczyk, J., Sobkow, A., and Zaleskiewicz, T. (2015). Affect-laden imagery and risk taking: the mediating role of stress and risk perception. *PLoS One* 10:e0122226. doi: 10.1371/journal.pone.0122226

Västfjäll, D., Slovic, P., Burns, W. J., Erlandsson, A., Koppel, L., Asutay, E., et al. (2016). The arithmetic of emotion: integration of incidental and integral affect in judgments and decisions. *Front. Psychol.* 7:325. doi: 10.3389/fpsyg.2016.00325

Vuilleumier, P., Armony, J. L., Driver, J., and Dolan, R. J. (2001). Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron* 30, 829–841. doi: 10.1016/S0896-6273(01)00328-2

Vuilleumier, P., and Schwartz, S. (2001). Emotional facial expressions capture attention. *Neurology* 56, 153–158. doi: 10.1212/WNL.56.2.153

Watson, D., and Clark, L. A. (1999). The PANAS-X: manual for the positive and negative affect schedule - expanded form. *Iowa Res. Online* 277, 1–27. doi: 10.1111/j.1742-4658.2010.07754.x

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., and Peters, E. (2013). Development and testing of an abbreviated numeracy scale: a rasch analysis approach. *J. Behav. Decis. Mak.* 26, 198–212. doi: 10.1002/bdm.1751

Wichary, S., Mata, R., and Rieskamp, J. (2016). Probabilistic inferences under emotional stress: how arousal affects decision processes. *J. Behav. Decis. Mak.* 29, 525–538. doi: 10.1002/bdm.1896

Wulff, D. U., Mergenthaler-Canseco, M., and Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychol. Bull.* 144, 140–176. doi: 10.1037/bul0000115

Zaleskiewicz, T., Gasiorowska, A., Kesebir, P., Luszczynska, A., and Pyszczynski, T. (2013). Money and the fear of death: the symbolic power of money as an existential anxiety buffer. *J. Econ. Psychol.* 36, 55–67. doi: 10.1016/j.joep.2013.02.008

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

HYPOTHESIS AND THEORY
published: 31 October 2018
doi: 10.3389/fpsyg.2018.02073

Decisional Dimensions in Expert Witness Testimony – A Structural Analysis

Alex Biedermann[1,2]* *and Kyriakos N. Kotsoglou*[3]

[1] School of Criminal Justice, Faculty of Law, Criminal Justice and Public Administration, University of Lausanne, Lausanne, Switzerland, [2] Litigation Law Unit, University of Adelaide Law School, Adelaide, SA, Australia, [3] School of Law, Liverpool Hope University, Liverpool, United Kingdom

The relationship between forensic science and legal adjudication is intricate mainly because the need to inform fact-finders on issues going beyond the layman's knowledge poses challenges both on empirical and normative dimensions, in particular with regards to the specific role and duties of the different participants in the legal process. While rationality is widely upheld as one of the aspirations of the legal process across many modern jurisdictions, a pending question is how to remedy the uneasy relationship between general propositions (and knowledge claims) conditioning expert witness testimony, and individualized decisions taken by fact-finders. The focus has hitherto been put on the utilization of model-based and formal methods of reasoning while, regrettably, the concepts of judgment and decision-making have not received equal attention. A first aspiration of our paper will thus be to further clarify the nature of this systemic relationship in the particular area of the legal process involving scientific experts, by conducting a critical transversal analysis of current empirical, normative and doctrinal understandings of expert witness testimony. As a second aim, we will use this insight to argue in favor of the view that structural features of expert witness testimony are embedded in a decision-making process, and that the understanding of this decisional dimension is important for clarifying the respective roles of expert witnesses and fact-finders, and for favoring their mutual understanding thereof. To substantiate this perspective, and attest to its growing recognition as a frontier understanding, we will provide real-world examples from forensic science reporting practice and policy documents of professional bodies.

Keywords: expert evidence, legal process, decision analysis, normative approach, decision-making prerogative, expert witness fallacy

> "[Y]our degree of belief does not, by itself, dictate what you should *say* or *do* (. . .) A rational decision about what to do requires more than the evidence you have" (Sober, 2008, at p. 7)

OPEN ACCESS

Edited by:
Nathan Dieckmann, Oregon Health & Science University, United States

Reviewed by:
Christian Dahlman, Lund University, Sweden
Michael Stanley Moss, Northumbria University, United Kingdom

***Correspondence:**
Alex Biedermann
alex.biedermann@unil.ch

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

Received: *24 April 2018*
Accepted: *09 October 2018*
Published: *31 October 2018*

Citation:
Biedermann A and Kotsoglou KN (2018) Decisional Dimensions in Expert Witness Testimony – A Structural Analysis. Front. Psychol. 9:2073. doi: 10.3389/fpsyg.2018.02073

INTRODUCTION

In law, as much as in other disciplines, the topics of judgment and decision making (JDM) under uncertainty have both a long-standing and lively debated history, on all common levels of consideration, normative[1], descriptive and prescriptive. Qualitative verbal decision criteria, such as 'beyond reasonable doubt' (BRD), are typical examples that lend themselves for study under these distinct perspectives. Most textbooks in the field present standard legal decision criteria, but the competing interpretations of the nature and logical structure of such terms divides practitioners and scholars since decades. Normativists – to name one of the groups of discussants– have analyzed and expressed legal standards of decision in terms of formal frameworks, such as (Bayesian) decision theory, since the 1960s (i.e., Kaplan, 1968). Their research led to interesting analytical results, some of which confirm the meaningfulness of conventional decision standards, such as the $>50\%$ probability requirement for finding civil liability (e.g., Friedman, 1997; Kaye, 1999). In such normative frameworks, minimal probability thresholds required to justify particular decisions are tightly bound to preferences among decision consequences through so-called consistency relationships (e.g., Buchak, 2016). Formally, these relationships amount to principles such as minimizing (or maximizing) expected loss (or utilities). Substantial amount of research (for a review see, e.g., Connolly, 1987) has been devoted to the empirical investigation of the extent to which people's actual thinking and deciding in legal applications aligns to such consistency relationships between, on the one hand, beliefs about competing versions of the event of interest (i.e., hypotheses or propositions), expressed in terms of probabilities, and on the other hand preferences among decision outcomes[2], expressed in terms of utilities or losses. Such empirical work found that there is a considerable mismatch between the decisions that individuals were willing to make, that is decision behavior, and the decisions that would be optimal according to the normative account.

Further investigation of these results seems to have come to halt because of the exhausted perspectives that they represent. Normativists, for example, argue that precepts following from logical considerations simply are not and cannot be invalidated in principle by any mismatch with practically observable decision behavior that there may be. According to this view, the poor mapping of formal theories on legal adjudication does not represent a failure, because description and explanation is not one of the aspirations of these theories. Empiricists, in turn, consider aspects of normative perspectives pointless because of the absence of legal requirements in the first place that would ask participants in the legal process to conform to aspects such as the maximization or minimization of an expectation of any quantity of interest, such as utility or loss (e.g., Allen, 2003). This suggests that there is an impasse between the reality faced by legal practitioners and the many conceptual accounts offered by decade-long legal research and scholarship in formal approaches to inference and proof, including empirical studies by experimental psychologists. Hence, contributing to a collection in the area of judgment and decision making in this journal of applied psychology poses not only a high burden of providing original discussion. It also requires a clear statement of the scope of enquiry within the broad perspectives of descriptive, normative and prescriptive research.

We address this challenge by focusing our attention not to the process of legal adjudication as a whole, in particular ultimate issues to which most of the abovementioned decision-theoretic research relates, but to the intersection between forensic science and legal adjudication, in particular the form, content and elicitation of forensic expert conclusions. The reason for this is that, first, while debates over the appropriate approach to the various dimensions in which legal adjudication seeks optimization[3] seem stalled, there are local instances of the legal process, such as the use of specialized (forensic science) evidence, that represent unresolved conceptual difficulties in practical proceedings. Some of these difficulties are peculiar to the intersection between science and the law, such as the deferential versus educational approaches to deal with specialized knowledge in the process (e.g., Miller and Allen, 1993). A second reason is that devising a coherent approach to such challenges is an important preliminary to sound decisions at higher levels in the process, of which forensic expertise may be an integral part. By forensic evidence we mean, throughout this paper, both physical/chemical and digital *non-replicable* items of evidence, usable to help recipients of expert information discriminate between competing propositions of interest, or help reduce the pool of potential persons/objects at the origin of a particular trace or item seized in relation to an event of legal interest (criminal, civil, or administrative). Examples for particular types of forensic traces are mentioned in Section "Decision-Structures for Specific Evidence in Forensic Science."

In essence, our analysis will come down to, and articulates what we will call 'decision-structures.' We show that decision-structures, although normative in nature, conceptualize and ascribe content to existing adjudicative practice, for they capture the requirement of 'specific evidence.' Central to this argument is that the proposed decisional perspective is not an end in itself, but only a necessary preliminary to understanding the reason for and justification of counter-current positions, such as the call to abandon some traditional expert reporting formats; especially categorical conclusions that usurp the epistemic rights of fact-finders. This result will call into question the extent and scope of some of the current and most longstanding forensic science reporting schemes. To redirect such forensic

[1]In this paper, we understand the term 'normative' not in a legal sense (i.e., referring to a legal norm or precept), but in the way commonly understood in the JDM literature of applied psychology, that is as a logical standard against which people's reasoning and decision making can be evaluated (e.g., Baron, 2008, 2012). See also Hahn (2014) and Oaksford (2014) for accounts of normativism in the JDM literature.

[2]By decision outcome we mean a decision taken in the light of a particular state of nature. For example, the decision to convict a truly liable defendant is an accurate decision outcome, whereas a guilty verdict for a truly innocent defendant is an inaccurate decision outcome (i.e., a false conviction).

[3]See, e.g., Allen (2015) for an account on dimensions such as the organization of trials, governance, social concerns, and enforcement issues.

testimony on its proper track, recipients of expert information need to assume a more active role in the processing of scientific evidence, by insisting on their role as ultimate arbiters of probative value in criminal trials, in particular by explicating their exclusive epistemic duty to reach contextually structured decisions.

Methodologically, we will rely on the view according to which the logical and balanced assessment of scientific evidence is a central part of forensic expert testimony. Though, traditionally, this is said to involve probability as a measure of uncertainty (e.g., Aitken et al., 2010), we will adopt a broader perspective here and consider forensic expert testimony as an instance of a normatively structured decision-making process under uncertainty. Thus, we regard forensic expert testimony not as an abstract and isolated object of inquiry but blend it with considerations from actual forensic practice (e.g., policy documents and practitioner guidelines). This will also prompt us to assess the ways in which this perspective may contribute to the improvement of frontier understandings about the processing of scientific evidence in legal adjudication.

Our paper is structured as follows. Section "Transversal Overview of Current Empirical, Normative and Doctrinal Understandings of Expert Witness Testimony" critically reviews current perspectives on expert witness testimony. Using practical examples, we will expose areas of interaction between forensic science and the law where conflicting views about the form and the content of expert testimony continue to pose challenges for the legal resolution of disputes. Based on this initial diagnosis we will argue, in Section "Decisional Dimensions of Forensic Expert Testimony," that considering expert testimony not merely as an inference problem, but analyzed as a contribution to a process of decision, dissolves key aspects of current controversies without breach with either logical considerations or procedural principles. Discussion and conclusions are presented in the last section.

TRANSVERSAL OVERVIEW OF CURRENT EMPIRICAL, NORMATIVE AND DOCTRINAL UNDERSTANDINGS OF EXPERT WITNESS TESTIMONY

Key Controversies Over Selected Aspects of Forensic Expert Testimony

Traditionally, forensic science is regarded as a collection of applied scientific methods and techniques for the purpose of assisting the judiciary in specialized matters where it lacks relevant knowledge and expertise. While science and technology are subject to continuous change and development, conceptual questions gravitating around the quantification and weighing of scientific findings tend to concentrate on a singular, well settled perspective. Evett (2009, p. 159, emphasis as in original) has expressed this perspective as follows: "the single most important advance has nothing to do with technology [...]. It tells us the most important lesson for the logic of evaluative forensic science: *consider the probability of the evidence, given the proposition.*" How could such a simple sentence – at least at first sight – be considered the most important lesson for evaluative forensic science? A main reason is that it clearly delineates the area of competence of the expert, as noted by Margot: "[w]hether these results [are] observed if one proposition for the event is true rather than another proposition is the central relevant matter on which the forensic scientist may comment" (Margot, 2011, p. 796). This focus is fundamentally different from that of fact-finders who concentrate on propositions, given the evidence (Robertson et al., 2016). Yet these different logical conditionings, in particular evidence *given* target propositions (and the reverse), trouble both scientists and recipients of expert information since decades (Thompson and Schumann, 1987): it is the archetype forensic science and legal adjudication example for a normatively sound principle that is practically poorly understood. Many past and recent initiatives, including efforts by renowned scientific societies (e.g., Aitken et al., 2010), focus on explaining and exemplifying these principles through guidelines, recommendations and primer documents (e.g., The Council of the Inns of Court [COIC] and The Royal Statistical Society [RSS], 2017).

Often, however, the above state-of-the-art occupies only a side-arena of broader debates over forensic conclusion formats, with different discussants pulling the rope in different directions. Proponents in forensic fields that pursue the idea of identification (also sometimes called 'individualization'), provide typical examples for this. Identification, in the present context, is widely understood as the reduction of a pool of potential sources (of a crime stain, mark or trace) to one and only one single candidate (i.e., a person, object or tool). Examples of traces are biological stains (e.g., blood, saliva, etc.), marks on fired bullets, bite-marks, handwriting/signatures etc. and examples of conclusions are 'this DNA comes from *this* person,' or '*this* mark comes from this tool/person' etc. The unscientific character of such categorical conclusions (i.e., certainty assertions) has been prominently exposed by Stoney (1991) in his landmark paper "What made us ever think we could individualize using statistics?", but remains widely unrecognized. Not only are identification/individualization conclusions by scientists logically untenable, it has also been shown empirically that forensic examiners, in many instances, cannot make such determinations reliably, or at least exhibit a potential of error. This has encouraged calls to initiate a paradigm shift (e.g., Saks and Koehler, 2005), but the effect merely was to keep the topic on the agenda, leaving fundamental changes by practitioners pending, even in the light of subsequent, critical reports by the National Research Council [NAS] (2009) and, more recently, the President's Council of Advisors on Science and Technology [PCAST] (2016).

This divide over forensic reporting formats also surfaces on institutional levels, revealing the profound gaps between legal and scientific proponents. Most recently, the Department of Justice released a document entitled "Approved uniform language for testimony and reports for the forensic latent discipline" (U.S. Department of Justice [DOJ], 2018) which, contrary to the above considerations, upholds bold statements

such as "identification" and "exclusion".[4,5] This does not fit well with the release, in 2016, by the Office of the Attorney General (U.S. Department of Justice), of a memorandum to advise *against* categorical conclusions (certainty conclusions of the type mentioned above): "Department forensic laboratories will review their policies and procedures to ensure that forensic examiners are not using the expressions "reasonable scientific certainty" or "reasonable [forensic discipline] certainty" in their reports or testimony. Department prosecutors will abstain from use of these expressions when presenting forensic reports or questioning forensic experts in court unless required by a judge or applicable law" (Office of the Attorney General, U. S. Department of Justice [OAG], 2016). While this seems to be a clear message, the position of scientists remains far from uniform. For example, in a position statement regarding the OAG memorandum, some practicing DNA scientists re-asserted their adherence to categorical reporting formats (i.e., identifications), called 'source attribution determinations' (e.g., Moretti and Budowle, 2017). Unstated, however, remains the fact that this statement is based on particular assumptions that, in operational casework, are highly debatable (e.g., the omission of the potential of error), thus making the position both peculiar and difficult to defend. More generally, positions of this kind represent only one side of the extremes that characterize the concurrent streams of development in forensic science. Suffice to notice, as a counter example, that Cole (2014, p. 144) concluded, in a meticulous review of forensic (fingerprint) analysis practice, that "forensic identification will have difficulty moving forward until 'individualization' is really dead." Such stark language also continues to emanate from most recent discussions. A concise illustration for this is given by Evett who, at a NIST colloquium in 2017, has been quoted as saying "The identification paradigm is going to die, because as scientists we realize there's no basis for it" (Champod and Evett, 2017).

Despite these fundamentally opposed views on forensic expert conclusions, there is one common thread to which all discussants appear to subscribe: the idea of contributing to sound decision-making. Yet, strangely, current discussions in both practice and literature almost exclusively focus on questions about the nature, foundations and internal consistency of expert witness testimony, leaving aside the crucial question on how testimony interfaces with decision-making in the wider, albeit structured and detailed context of legal adjudication. We will critically expose this interface in further detail in Section "Decisional Dimensions of Forensic Expert Testimony," introducing the notions of decisional dimension and forensic decision structures. It is first necessary, however, to introduce elements from law (see Section "Law of Evidence, Complexity and Decision-Making Prerogatives"), in particular evidence law, and considerations of how legal orders deal with science, as exemplified by landmark decisions such as *Daubert*, followed by its subsequent discussion by legal commentators (see Section "Law and Philosophy of Science"). These preliminaries aim at providing the wider scene wherein which the decisional dimension of expert witness testimony, presented in Section "Decisional Dimensions of Forensic Expert Testimony," is to be understood.

Law of Evidence, Complexity and Decision-Making Prerogatives

As a preliminary, it is important to recognize that specialized forms of evidence, such as forensic science evidence, are merely instances of the broader challenge of evidence processing. Notwithstanding, the intersection between forensic science and legal (especially criminal) adjudication is often considered a prime example for illustrating the 'problem' of specialized knowledge, generally termed expert evidence (or, expert witness testimony) throughout this paper. It is commonly understood that the need to inform fact-finders on issues going beyond the layman's knowledge (see Section "Key Controversies Over Selected Aspects of Forensic Expert Testimony") poses challenges both on empirical and normative dimensions, in particular with regard to the specific role and epistemic duties and rights of the different participants in the legal process.

The very possibility of accurate and efficient legal operations hinges on the ability of fact-finders to recognize particular acts and circumstances as instantiations of abstract legal concepts. It is widely recognized that this capacity has reached new limits in today's technology-driven modern world, with its wide range of socioeconomic activities. What is more, it is questionable whether laypeople can appropriately comprehend evidential items or phenomena and assess their informative contents with respect to the contested facts of the case, when this requires trained expert sensory capacities and specialized knowledge. The technological advances of our age raise thus pressing questions of competence: who should have the *decision-making prerogative* regarding selected conclusions (e.g., regarding the origin of a particular mark or stain, item of handwriting, etc.) when dealing with items of evidence that require knowledge fact-finders do not usually possess? This central issue will be addressed in later Sections of this paper.

Interestingly, from a historic point of view, this is not the first time that the increasing complexity of society and compartmentalization of human knowledge place additional strain on legal systems and the "good old way," according to which expert witnesses act simply as "helpers of the court" (Thayer, 1892, p. 665). Legal history gives relevant insight into the dynamics of the law of evidence (Golan, 1999). During the Middle Ages, jurors in predominantly agricultural societies – whose level of sophistication and technological advance was not radically different from that of ancient communities – were drawn from the immediate surroundings of the accused (Langbein, 1996, p. 1170). The rationale underpinning this adjudicative structure was the assumption that the jurors would either be familiar with the allegation at play or they would be able to investigate on their own. As small communities gave their place to increasingly larger ones, the institution of self-informing jurors underwent fundamental changes, for the law

[4]Note that similar documents exist in other jurisdictions. An example is the German 'Standard of fingerprint identification' [translation by the authors; original title: 'Standard des daktyloskopischen Identitätsnachweises'] [BKA (Bundeskriminalamt), 2010].

[5]For a timely and critical review, see Cole (2018).

covered gradually broader areas of social life. There was simply too much information to navigate and it was now passive fact-finders, ignorant of the contested facts and essentially dependent on witness testimony, who decided cases. These laypeople with no direct knowledge of the facts were, obviously, in need of judicial instructions regulating the routines for gathering and assessing evidence, i.e., information, in order to render a verdict. According to John Henry Wigmore, the law of evidence grew, at this very moment, as procedural necessity and doctrinal reality. It was, in other words, the dawning of the instructional trial propelled by the need to acquire (specialized) information that molded the law of evidence (Wigmore, 1908, p. 692).

The complexity of legal proof seems to swing the pendulum back in the direction of active rather than passive fact-finders, suggesting that they are wearing this time the hat of 'expert witnesses.' For the democratically legitimized and from the legal order authorized professional judges or jury cannot be the ones, so a general claim in forensic science, who make so-called identification decisions (as defined in Section "Key Controversies Over Selected Aspects of Forensic Expert Testimony"). According to the deference model, fact-finders will have – on the pain of irrationality – to delegate some of their cognitive monopoly to experts, at least every time the contested facts feature questions encroaching beyond the boundaries of what is commonly known.

On a practical account, however, the notions of decision-making prerogatives and deference lead to a critical impasse. On the one hand, the procedural necessity of filling abstract legal terms with valid (and reliable) empirical content highlights the systemic relationship between legal adjudication, especially criminal adjudication, and (forensic) science. For one of the central tenets of modern legal orders, the Rationalist Tradition, is the requirement that all decisions, which affect the interests of individuals by resolving disputed questions of fact, are justifiable (Twining, 1982). A decision-making process in which the fact-finder does *not* properly understand the nature (e.g., statistical) and empirical content of evidence would be arbitrary and have deleterious effects for the public confidence in the integrity and accuracy of the legal system. Lord Steyn's dictum that "[c]ourts of law can only act on the best scientific understanding of the day"[6] entails the admission that fact-finding can be as good as modern science allows it to be. On the other hand, the relationship between the two symbiotic partners is characterized by friction and antagonism. Forensic scientists take the legal axiom 'iudex non calculat' quite literally and deplore that traditional fact-finders (judges and jurors) struggle with the proper understanding of scientific methods, and science in general. At the same time, the state of forensic science causes itself, in regular intervals, scorn, and even ridicule.[7] What is more, in her recent annual report, the U.K. Forensic Science Regulator concluded that failing forensic science standards make "miscarriages of justice inevitable."[8]

The intermediate interrogation at this point thus is how to move on from this impasse. In order to reflect further on this systemic relationship and assist in finding ways to cope with its difficulties, it is necessary to lay down some basic rules of conduct. Articulating the structural features of a normative platform favoring communication and mutual understanding between juridical decision-makers and expert witnesses can be seen as a mapping exercise, aiming at drawing normative borders and allocating epistemic rights and duties. Between fact-finders' reaching beyond their legitimate scope of their expertise on the one hand and expert witnesses trespassing on the realm of the jury on the other hand, the challenge is to strike a scientifically defensible and jurisprudentially fair balance. But in order to avoid a mapping exercise to fall short of practical considerations, it is necessary to take into account the architecture of adversarial criminal adjudication,[9] policy choices and methodological axioms of science. What is more, rethinking expert witness testimony has to take place during business-as-usual operation, avoiding interference with the established routines for generating, evaluating and validating knowledge claims in legal adjudication. Adjusting the normative structure of legal institutions to theoretically sophisticated models does not mean that one is authorized to change the structure of dispute resolution in autonomous legal orders. This perspective is not limited to criminal adjudication as forensic evidence can also play a crucial role in civil lawsuits (see, e.g., forensic document examination in the case *Zuckerberg v. Ceglia*).

Law and Philosophy of Science

An analysis and discussion of forensic expert testimony is hardly possible without devoting some comments to the relationship between law and philosophy of science. The interest of the former in theories in general, and in questions such 'what is science?', is not restricted to academic circles. In fact, systems of legal adjudication are pragmatically – and in terms of substantive law – required to filter the admission of theories in their proceedings. This is a relevant observation because practitioners and forensic scientists will base propositions of interest, to some extent, on theoretical models. The principal issue with this is that in most jurisdictions it is the judge rather than the respective scientific community that will have to answer the question of (scientific) validity. Thus, introducing elements of philosophy of science at this juncture aims at further delimiting the focus of enquiry and clarify the nature and the scope of the decisional account introduced later in Section "Decisional Dimensions of Forensic Expert Testimony."

[6] *R v Ireland; R v Burstow* [1997] 3 WLR 534, House of Lords, per Lord Steyn.

[7] See e.g. the recent negative publicity blitz strike at Jon Oliver's political satire show on the shortcomings of forensic science, Last Sunday Tonight, Season 4, Episode 25 (2017).

[8] See https://www.theguardian.com/uk-news/2018/jan/19/uk-police-forces-failing-to-meet-forensic-standards-safe-regulator-miscarriages-justice-outsourcing (01.25.2018).

[9] We do not wish to make any specific claim on the differences between adversarial and continental systems of adjudication, given that from our structural point of view, differences tend to disappear. For reasons of simplicity and in order to provide concrete and comprehensible examples we will focus on adversarial systems of criminal adjudication. See Damaska (1973) for more discussion.

The engagement of courts with philosophy of science has been hitherto rather light-hearted (Haack, 2014a, p. 141). But the academic discussion too, had difficulty to comprehend the scientific endeavor of expert witnesses against the background of the model-based view of scientific enterprise. The U.S. Supreme Court's ruling on *Daubert*[10] is particularly pertinent in this context as it has become a leading authority on monitoring the reliability and validity of expert evidence in all US Federal Courts and the majority of state jurisdictions. Most importantly, it has generated a remarkable amount of academic discussion on an international level, despite it not being directly relevant for proceedings in other jurisdictions. In *Daubert*, the petitioners (two minor children and their parents) had alleged that the children's serious birth defects had been caused by a prescription drug marketed by the respondent. They also proposed to adduce evidence of the testimony of eight experts to the effect that the prescription drug can cause such side effects. Both the District Court and the Court of Appeals relied on Frye's standard of admissibility,[11] i.e., general acceptance, and declared the evidence inadmissible. For there was extensive published scientific literature on the subject that the maternal use of Bendectin has not been shown to be a risk factor for human birth defects.

The US Supreme Court declared that the rule of 'general acceptance' had been invalidated by the adoption of the Federal Rules of Evidence. The ruling signified departure from *Frye* and reliance on "general acceptability," for the *Frye* test was superseded by the Rules' adoption, in favor of a more liberal approach. According to Rule 703, Justice Blackmun who delivered the opinion of the Court said, the question that is most pertinent for the court is whether "all scientific testimony or evidence admitted is not only relevant, but reliable."[12] Let us stress that *Daubert* was about admissibility of the evidence, not weight. Undeniably, admissible evidence does not predetermine decisions. It changes, however, the dynamics of proof. The Court explicitly coupled evidentiary relevance qua precondition of admissibility and scientific validity. While the Court was at pains to stress that the central issue is validity, not a specific criterion thereof, it introduced a "flexible" inquiry encompassing multiple and non-exhaustive factors, which, the Court reminds us, do not "set out a definitive checklist or test."[13] Criteria such as whether the theory or technique underpinning the evidence has undergone testing and withstood the scientific process of falsifiability; whether it has been subjected to peer review and publication in refereed journals; whether there is information about its known or potential error rate; whether the theory or technique enjoys the support of some relevant scientific community or communities.[14]

Daubert has attracted wide criticism especially with regard to the Popperean criterion of falsifiability with authors going at great length to flesh out the philosophical argument. Allen (1994, p. 1164) concisely remarked that the Court "replaced a judicial anachronism [*Frye*] with a philosophical one [Popper]."[15] Haack remarked, in the same tone, that Daubert's "philosophy of science was confused" (Haack, 2014b, p. 113). To some extent, the unusual amount of criticism against *Daubert* overstates, in our opinion, the importance of a single member of a non-exhaustive set of criteria of validity, i.e., falsifiability. The inquiry enshrined in Rule 702, Justice Blackmun clarifies, is a flexible one and its overarching subject, to wit, scientific validity, cannot be reduced to any single criterion.[16] Falsifiability, peer-review process etc. are simply indicators, not necessary and sufficient conditions of scientific validity. The major change that *Daubert* engineered is the shift from an externalist approach to scientific evidence to an internalist one. Whereas for 70 years judges would have to rely on the "general acceptance test," they now need to comprehend the empirical claims and underlying methodology, for admissibility hinges on *asserted* scientific validity. It does not follow, thus, that the Court subscribed to Popper's conception of science, for the Court did not answer authoritatively the question of what constitutes 'validity.' Secondly, *Daubert* is important for the contradistinction between science and legal adjudication. While "[s]cientific conclusions are subject to perpetual revision," the Court points out, "law . . . must resolve disputes finally and quickly."[17] This remark raises more questions than it manages to answer. For one, it seems to assume that science is on a further end of some spectrum, investing more time and resources than legal adjudication does. If this is true and we can directly compare the two systems, one might wonder, then how are we to avoid the conclusion that scientific methods *are* superior to adjudicative ones? How could we justify the limited role expert witnesses are required to play, when modern legal orders equate parts of their testimony with trespassing on the province of the jury? To tackle these questions in the context of legitimacy of criminal adjudication, it is necessary to take a closer look at the philosophy of science encoded in *Daubert*.

Justice Blackmun seems thus to have placed some emphasis on the criterion of falsifiability, and has attracted waves of criticism ever since. A key question, he writes, in determining whether a theory or technique is scientific knowledge and therefore reliable and admissible in court, is whether it "can be falsified."[18] The Popperean scent ascending from the criterion of falsification is enhanced, for Justice Blackmun uses a direct quotation in his next sentence: "the criterion of the scientific status of a theory is its falsifiability, or refutability, or testability."[19] While Popper is a widely respected philosopher, especially among lawyers, his views have, however, never had an actual impact on existing and established methods for validation of scientific hypotheses. As Kuhn (1996, p. 77) remarked, "[n]o process yet disclosed by the historical study of scientific development at all resembles the

[10] *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993).

[11] *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923). According to the old Frye Rule, novel scientific testimony needed to be generally accepted in the relevant field in order to be declared admissible.

[12] *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993), at 589.

[13] *Id.*, at 593.

[14] *Id.*, at 580.

[15] Allen is in turn citing one of his students.

[16] *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993), at 594.

[17] *Id.*, at 597.

[18] Id., at 593.

[19] Id., at 593; Popper (1962, p. 37).

methodological stereotype of falsification by direct comparison with nature."

But the criterion of falsification was not the only hint to philosophy of science in *Daubert*. In the same paragraph Justice Blackmun mentioned also Carl Hempel, a central figure of Logical Positivism. This has been criticized as cherry-picking of philosophical ideas (Haack, 2014a). On the one hand, the criticism is justified, for there are major differences between Hempel's verificationism and Popper's falsificationism. On the other hand, the Supreme Court hit the nail on the head, apparently without even realizing it, because both meta-theoretical approaches share at least two structural features. Firstly, the presupposition that theories are based on some formal logical syntax, to wit, by carrying out axiomatization of theories within formal languages. Secondly, the idea that a scientific theory could be once and for all confirmed or falsified through a direct comparison with a theory-external criterion, i.e., a theory-neutral observational language. However, phenomenal appearances can only be validated in the light of a multitude of background assumptions (Jackson, 1988, p. 557). The theory ladenness of all experiential data, i.e., one of the radical insights of the second half of the 20th century, necessitated the abandonment of the strict divide between theoretical terms and observational ones. All in all, despite their differences, these two approaches, i.e., Hempel's verificationism and Popper's falsificationism, can be regarded from the stance of philosophy of science as the two main phases of the *syntactic view* of theories which dominated the first half of the previous century. The *syntactic view of theories* (Received View) with its phantasies of an ideal language comprising concepts with sharp boundaries, which dominated the field of philosophy until the 1950s, and its underlying logicism overestimated the power of formal logic. What is more, it failed to give a practicable account of actual and successful scientific theories (Suppe, 2000, p. S103). The latter are not axiomatic systems yielding deductively derived consequences, and scientific practitioners have more moderate requirements for scientific validity than a "relentless accumulation of confirming instances" (Toulmin, 1967, pp. 110–111). Scientific propositions are not based on strict unexceptional 'laws,' but on *generalizations*. The *semantic view* of theories, in turn, which has been dominant since the last quarter of the 20th century, is a formal reaction to the syntactic view of theories (Bailer-Jones, 2009, p. 126). It is remarkable, that both Courts and the academic discussion appear to have largely failed to register this major development, i.e., the fact that "[m]odels occupy central stage" (van Fraassen, 1980, p. 44) in philosophy of science.

Let us synthesize the above. According to the model-based view of scientific enterprise, theories are not empirically un-interpreted formal-axiomatic systems but involve a central interpretative aspect. In other words: theories are not fully axiomatized systems which eliminate the need for discretion in science let alone in legal adjudication. This highlights the need to outline the area of admissible interpretation for expert witnesses and fact-finders. Further, scientific models do not consist in the accumulation of instances who either confirm or fail to falsify a given hypothesis once and for all. An essential feature of modeling is to generalize. Generality, understood as the property of applying widely, plays a pivotal role in philosophy of science (Lewis and Belanger, 2015). Furthermore, it is important to understand the inversely proportional character of generality and precision. As Gleick (1998, p. 278) points out: "The choice is always the same. You can make your model more complex and more faithful to reality, or you can make it simpler and easier to handle." This trade-off, however, is, no matter the outcome, subject to certain restrictions. The purpose of any model is to generalize and reduce reality to meaningful theoretical (i.e., general) propositions. A map which would be as accurate as the landscape itself would be a contradiction in terms. We can hold, therefore, that there is a point where any general account of the world breaks down. That point is the individual case. This insight, which is methodologically rather trivial but as regards its consequences radical, helps us realize the different dynamics and aspirations between legal adjudication and scientific endeavor.

DECISIONAL DIMENSIONS OF FORENSIC EXPERT TESTIMONY

Discretion in Law

Legal Conclusions and Decisions Versus Scientific Determinism: The Need for Discretion

As argued in Section "Transversal Overview of Current Empirical, Normative and Doctrinal Understandings of Expert Witness Testimony," the function of any model providing scientific explanation is (i) to generate generalizable propositions (conclusions), presuming that "events occur in consistent patterns" (National Research Council [NAS], 2009, p. 111), (ii) to establish symmetry across members of a target system, and (iii) to eliminate the need for case-by-case treatment of individual cases. The validity of a general proposition, i.e., its scientific character, is a function of its derivability from a scientific model to such an extent that the very expression '*ad hoc* explanation' strikes us as quite peculiar, indeed as a contradiction in terms. Singularities, where physical laws break down, are deeply troublesome for scientific theories. The question, then, is whether the fact-finder's decision about unique historical events is also generalizable. From Aristotle, who observed that it is "foolish to accept probable reasoning from a mathematician and to demand from a rhetorician [i.e., lawyer] scientific proof,"[20] to modern forensic scientists who are at pains to stress that the idea of a frequency being attached to an outcome for a single event is "ridiculous" (Lucy, 2006, p. 5), scholars have continuously rejected (bogus) claims of generality when it comes to legal decisions. However, we will need more than aphorisms, in order to draw a line between fact-finding in legal adjudication and scientific inquiry.

The fact that legal systems, especially in our increasingly complex world, are unable to, indeed not particularly interested in predicting and axiomatizing every combinatorial possibility of circumstances that the future may bring – this would be computationally intractable – is an enduring lesson we have learnt from the failures of legal orders that placed exclusive emphasis on casuistry and tried to provide an all-encompassing

[20] Aristotle Nic Ethics i3, 1094b.

solution to the problem of decidability by enacting exhaustive lists of elements falling under a legal concept 'φ'.[21] The (vain) effort to provide an ontological map of semantics of legal concepts aimed at the elimination of discretion and predicated a fact-finder/judge who would effectively be 'the mouth of the law.' However, this presupposes a world comprising a finite number of features, so that we could lay down rules for each combination individually. As Hart put it, "[p]lainly this world is not our world" (Hart, 1961, p. 128). Explication of legal terms, including "proof," is highly contextual. Securing a maximum degree of predictability (and therefore: legal certainty) comes at the price of "freezing" the meaning of legal terms by settling in advance issues before they arise. This would provide maximum legal certainty in lieu of paradoxical results, such as for example the prohibition of an ambulance entering a park, pursuant to the rule "*No* vehicles in the park." The rigidity of various legal classifications, e.g. what constitutes 'proof,' would make legal orders instantly obsolete and unfit for resolving new questions that will inevitably emerge in litigation. In a world characterized by a radically unpredictable future, every deterministic approach to legal concepts – let us mention again that from the point of view of the law, 'proof' is a legal concept – would be in need of revision moments after its enactment in order to catch the multitude of situations that occur in real life, and keep abreast of social developments. The whole field of legal methodology and legal dogmatics grew out of the ashes of legislative projects that tried to eliminate discretion only to fail utterly.

Modern legal orders have internalized the message that it is futile to anticipate decisions.[22] Admittedly, deduction from rules with predetermined meaning, elimination of discretion and the description of judge's/fact-finder's activity in logico-mechanical terms are features routinely attributed to juridical operations. The problem, however, is that these features derive from a rather superficial understanding of normative systems. Indeterminacy is not a surface feature of law but is inherent in natural languages. It is for this reason that logicians traditionally stress that truth values can be defined only within formal languages (Tarski, 1944). The dynamic process of increasing or decreasing the generality of legal rules inevitably runs into a point of bifurcation, where no decision either way is "dictated" by the applicable norm(s) (Hart, 1961). Notwithstanding the fact that the respective factfinder will probably have (good) reasons to reach the – in his or her opinion – right decision, from the point of view of the law there can only be a set of equally reasonable decisions. The applicable legal norm or standard in question is simply a "frame" within which various possibilities are given. The verdict, Kelsen explains, is on a thorough look an "individual norm," valid exclusively with regard to the individual case, i.e., not generalizable (Kelsen, 1934, para 36). At this very point, axiomatized systems break down, for the fact-finder needs to make a decision, which is not warranted by the underlying logical framework. Particular cases, Hart remarks, do not make themselves fit for legal subsumption, "already marked off from each other," or shouting at us: 'I am an instance of the general rule' (Hart, 1961, p. 126). Rules, including legal rules, do not provide the (meta-)rules for their own application.[23] The gap between rational *conclusions* based on scientific models and personal *decisions* about disputed facts can only be filled by an *act of will* (Kelsen, 1934, para 5), which is not a necessary outcome of a justificatory chain. We will come back to this important point.

It is worth keeping in mind that an uncontradicted model-based proposition can be rejected only on pain of irrationality. This is the essence of the deference model in forensic science. A fact-finder cannot simply disregard the justified *conclusion* (decision structure) of an expert witness testimony, say, on the assigned probative value of some biological trace. However, decisions behave in a different way. They are not rationally resolvable, for reasonable minds may differ. Disagreement about the 'one right decision' does not necessarily imply an error in the justificatory process, since the logical chain of justification leads to a point of progress branching with mutually incompatible growing paths which the decision-maker can follow (Stegmüller, 1979, p. 33). The fact-finder has the epistemic duty to exercise discretion and resolve an issue by making a decision. Scholars who deny this fundamental insight are obliged to postulate caricatures of judges with "superhuman intellectual power" (Dworkin, 1986, p. 239). What is more, discretion is not an exclusive feature of law. The historian of science Kuhn has promoted a similar view. He emphasized the role of value judgments and decisions in the course of scientific development. E.g., debates over theory-choice, he says, "cannot be cast in a form that fully resembles logical or mathematical proof" (Kuhn, 1996, p. 199). There is no algorithm, e.g. for choosing the level of significance (3σ or 5σ), in an experiment.[24] E.g., the existence of the Higgs particle is proved qua outcome of empirical research. Yet the underlying and staggering level of significance (5σ) is not itself a scientific fact; it is a convention and as such a matter of choice rather than of "purely theoretical reasons" (Stegmüller, 1979, p. 35).

The Values of (Criminal) Law

The previous considerations allow us to, first, articulate a principal source of confusion in discussions around expert witness testimony, and, secondly, explicate the decision-making prerogative alias *burden of decision*. Legal adjudication does not aim at, or aspire to answer empirical questions in a *general* way. It is not a shorter or less costly method of knowledge-claim validation. Its function and social task is to resolve, within a reasonable amount of time, a legal issue deriving from contested factual claims. Legal orders set general criteria, which, when met,

[21]The fact that legal systems operating in complex environments are unable to anticipate the future and contain rules allowing for exceptions incapable of exhaustive statements is a historic lesson we have learnt at least since the Prussian Legal Code (1794) with its more than 19,000 paragraphs.

[22]For example, the Criminal Law Revision Committee for England and Wales [CLRC] (1980) [Fourteenth Report, Offences against the Person (1980), Cmnd 7844, para 37] emphasized that they are "extremely hesitant about embodying in a statute (which is not always susceptible of speedy amendment) an expression of present medical opinion and knowledge derived from a field of science which is continually progressing and inevitably altering its opinions in the light of new information."

[23]Interestingly, the same holds for formal analytical frameworks, such as probability and decision theory.

[24]However, there is argument, at least in forensic science, to the effect that frequentist significance levels ought *not* to be used (e.g., Taroni et al., 2016).

authorize an official to impose a legal effect. The difficulty resides in the fact that the question *when* these general criteria ('exclusion of reasonable doubts' or 'being sure') are actually met, cannot be answered in the abstract. Social reality is complex and too context-sensitive for an algorithmic or axiomatized approach. Accordingly, the decision-making prerogative in actual cases refers to the responsibility to resolve an issue, not despite but because it is not replicable and therefore not subject to scientific analysis in the traditional sense. There is no univocal answer to the question of legal liability and proof because the question as such is not scientific, not because the underlying issue is obscure.[25]

Each legal order qua autonomous normative system will have to make a basic policy choice on who will take the responsibility and resolve a scientifically unresolvable – though scientifically describable – issue. The respective choice is not answerable to eternal and unalterable laws, but subject to historical contingencies, political balances and outcomes of social conflict. There is no *a priori* or scientifically valid reason to give the decision-making prerogative to professional judges, laypeople or experts, i.e., to opt for the educational or deference model. A decision is based on, albeit is not derivable by scientific propositions. It is pillared by the act of will of the respective official, who is authorized to make a decision, although no decision is logically necessitated by the underlying normative framework (Kelsen, 1934).

Utilizing an act of will does not mean that decision-making implies an anything-goes activity. Decisions are neither logically warranted, nor are they a step into the void. Each legal order has its own internal values, which the juridical decision-maker has to implement. The law especially criminal law and the criminal standard of proof are heavily influenced by policy considerations. The US Supreme Court has famously spelled out this dependency in the benchmark decision *In Re Winship*, which describes the reasonable doubt standard as "a prime instrument for reducing the risk of convictions resting on factual error."[26] The standard of proof (in a liberal legal order) reflects thus the increased social disutility of convicting a law-abiding citizen person. As Justice Harlan put it in his concurring opinion, the function of the standard of proof is to influence "the relative frequency of these two types of erroneous outcomes," knowing that the two types of error (acquitting the perpetrator and convicting the innocent) are inversely proportionate.[27] Similar considerations apply to almost every modern legal order.

Liberal legal orders, as opposed to authoritative ones, value the individualistic perspective, and the requirement that legal evidence has to be 'specific' cannot be sidestepped. As Justice Antonin Scalia put it, statistical evidence "is worlds away from [legally] 'significant proof'."[28] The idea that some scientifically validated (general) proposition guarantees the factual and normative rectitude of a verdict (decision) creates a "major contradiction between the scientific status that is claimed and the operational paradigm to which its practitioners subscribe" (Champod and Evett, 2001, p. 101). It is worth reminding that related discussion exists in the area of clinical decision-making, where it has become increasingly clear that despite common assumptions, 'diagnostic slam-dunks' and absolute certainty are the rare exception rather than the rule. Given the inevitable element of uncertainty in a typical diagnosis, the physician will be able to express, in a warranted way, merely the probabilistic support for some medical condition, e.g. tuberculosis, as compared to relevant alternative hypotheses. But the evidence alone, and the subsequent grade of belief, will *not* necessitate what should be done (Sober, 2008, pp. 4–5). The primary interest of a patient is the choice of therapeutic measures, not the probability of any disease. As Sober remarks, answering the question 'What should I do?' requires more than data and grades of belief. It requires the input of values (Sober, 2008, p. 4). The question whether the diagnosed condition corresponds with the true state of affairs and the related question, which treatment should be preferred, requires and instigates an inferential leap. Prominent forensic scientists call this step "a leap of faith" (Stoney, 1991, p. 198). (Forensic) scientists are, therefore, not better equipped than laypeople, to take this step by making a decision under uncertainty.

Forensic Reporting

Decision-Structures for Specific Evidence in Forensic Science

We can now exemplify our perspective on decision-structures by considering examples from forensic science reporting practice. We will focus on results of forensic DNA analyses that, despite critiques, are widely considered as a principal type of evidence, especially in criminal proceedings. The high variability of forensic DNA profiles between individuals has made it an attractive candidate for supporting claims of individualization. Traditionally, this goal has been conceived as the heart and soul of forensic science (Kirk, 1963, p. 236), and is also very common among other trace categories such as fingermarks, handwriting and the like, including also more recent trace types, such as digital traces.

Related to individualization is the notion of uniqueness which, however, is not an operable term in ordinary criminal adjudication. This hinges on methodological issues of the standard ways in which forensic scientists analyze biological traces. Forensic DNA profiling results reflect an individual's genetic features at various points of comparison, the so-called loci. But since only a tiny part of the entire DNA-molecule is analyzed, an eventual correspondence between the profile of a crime stain and that of a person of interest is, per definition only partial, and does *not* establish that the person of interest is the source of the crime stain (Redmayne, 1995, p. 464). This is especially the case for incomplete DNA traces (e.g., degraded trace material), or mixtures of DNA composed of material from more than one contributor. The probative value of DNA profiling results will thus be explicitly *probabilistic*, and it is essential to understand that probative value is based on the

[25] For a clarification of this point in the context of forensic identification see, for example, Biedermann et al. (2008, Section 5.2).

[26] *In re Winship*, 397 U.S. 358 (1970), Opinion of the Court (Brennan, J.), at 363.

[27] *Id.*, Harlan, J., Concurring Opinion, at 371–372.

[28] *Wal-Mart Stores, Inc. v. Dukes, et al.*, 564 U.S. 338 (2011), Opinion of the Court (Scalia), at 14.

notion of conditional genotype probabilities (hereinafter: CGP). The latter is a technical notion that expresses the probability of observing the DNA characteristics on the crime stain (i) given that an unknown person (i.e., different from the suspect or person of interest) is the source of the crime stain, (ii) given the task-relevant information available on the case file, and (iii) given additional considerations related to forensic genetic theory (Evett et al., 2000). The forensic biologist's assessment thus focuses on the probability of observing corresponding DNA characteristics in an unknown person from the relevant population (i.e., the population to which the source of the crime stain is thought to belong), which may be Caucasians, Chinese etc.[29] The probabilistic character of the respective report highlights the importance of distinguishing sharply between using a model to describe, *in a general way*, a phenomenon (i.e., the kind of genetic features observable on the crime stain) on the one hand, and using such information in order to make a legally structured and procedurally contextualized *decision* on the other hand. This categorical distinction has already been drawn by criminal courts.[30] Schematically, thus, two different questions regarding DNA evidence are commonly of interest:

(1) What is the probability that an individual will be observed to have the DNA profile features of interest as seen in the trace *given that this person was chosen at random from the population of interest*?
(2) What is the probability that a given individual is truly the offender (or, the source of the crime stain), *given that corresponding DNA features between the profile of that person and the profile of the crime stain have been reported*?

It is worth mentioning that in the *first* question, the factual innocence of the person of interest is taken for granted and one asks what the CGP is, whereas in the *second* question one takes the corresponding DNA features (i.e., the forensic findings) for granted and envisages the ascription of criminal liability (or, inference of source). As anticipated in Section "Key Controversies Over Selected Aspects of Forensic Expert Testimony," the widely known prosecutor's fallacy consists in transforming the answer to the first question into an answer of the second one. The prosecutor's fallacy is, however, not the only misinterpretation that one may make in relation to the above two questions. The further source of methodological error (and procedural violation) is what may be referred to as the 'expert witness's fallacy,' which refers to the situation in which forensic experts testify even further beyond the area of their expertise. This occurs, specifically, when experts purport to answer *both* questions, although they are legitimized to answer only the first. This sidesteps the understanding that "a sharp distinction can be made between what one ought to think about a proposition [. . .] and what one actually decides [. . .] the former is a problem that pertains to probabilistic reasoning whereas the latter is one that applies to decision making" (Biedermann et al., 2008, p. 23).

Interestingly, though not coincidentally, the above two separate questions map on the distinctive epistemic duties for expert witnesses and jury. The expert witness, the Court of Appeal in England and Wales makes clear, "should not be asked his opinion on the likelihood that it was the defendant who left the crime stain, nor when giving evidence should he use terminology which may lead the jury to believe that he is expressing such an opinion."[31] The expert witness, in other words, is logically warranted and legally authorized to express only the information regarding the probative value of the scientific findings (*not* an opinion regarding the truth or otherwise of the propositions of interest). Such information may take the form of, for example, the CGP, or a measure that is a function thereof. The situation is different, however, with the second question. Even in the factually remote, but epistemically possible case, where the correspondence would extend to more features than are included in traditional DNA profiles, and the source of the DNA is not contested by the adversarial parties, the question of liability would still not have been answered. Further considerations need to be taken into account for reasoning to such higher propositional levels, such as the relevance of the evidential material for the offense of interest (Stoney, 1994). In essence, associating a person of interest with evidential material, as such, does *not* answer any ultimate issue (i.e., a substantive element of an offense).[32]

This does not necessarily mean that the expert witness would be excluded in advance from answering non-scientific questions. The point merely is that there is no obvious reason to believe that expert witnesses would be any better than laypeople in answering questions such as individualization or liability. As Cole puts it, "the expert has no special competence greater than that of any other person at the decision stage of the process" (Cole, 2014, p. 143). Practically, one should even expect experts to be in a *less* informed position compared to the jury, because only the latter oversees the case as a whole. But again, it is the divide between reasoning about propositions, on the one hand, and actually making a decision regarding those propositions of interest, on the other hand, which poses both conceptual and procedural hurdles, and this clarifies why the expert is not in the position to act at the stage of decision. Guilt is not a (scientific) proposition, but a verdict, which is the result of a decision under uncertainty made after considering all elements of the case. The allocation of the decision-making prerogative is, intrinsically, a policy choice rather than a scientific mandate. Legal orders may choose freely whom they entrust with this important legal duty of deciding on the defendant's liability, without violating any logical or methodological principles of scientific inquiry.

Forensic scientists who – on an industrial scale – make categorical claims in terms of so-called individualization conclusions (see Section "Key Controversies Over Selected Aspects of Forensic Expert Testimony") with respect to the

[29]Such probabilities may become increasingly small, in particular smaller than one over several world populations, intriguing some commentators, including scientists, to assert individualization (i.e., uniquely assigning a person as the source of a given biological trace). However, such claims are unfounded because they take values produced by a biological *model*, operated at an extreme end of extrapolation and, thus, beyond what may be empirically investigated, as face values.

[30]*R. v. Deen*, The Times January 10.

[31]*R v Doheny and Adams* [1997] 1 Cr App R 369, 374.

[32]Rule 704(b) makes this point explicit when it states that "[i]n a criminal case, an expert witness must not state an opinion about whether the defendant did or did not have a mental state or condition that constitutes an element of the crime charged or of a defense. Those matters are for the trier of fact alone."

defendant, to the exclusion of all others, are wrong both in terms of methodological underpinnings and the law. They seem to conflate individualization qua ontological claim – according to which there is no other person who could be found to correspond to the *particular* DNA profile observed in the case at hand – and individualization qua epistemic claim leading to a definitive conclusion that has potentially decisive impact on the verdict (Saks and Koehler, 2008, pp. 211–219).[33] It is a longstanding, though still not widely appreciated, fundamental insight that for moving from the evidence to an individual as the proclaimed source of an item or trace, a leap of faith is required. To this we add, through our discourse here, that such forensic conclusions can also be framed as decisions, requiring an act of will (see Section "Decision-Structures for Specific Evidence in Forensic Science"). This adds further support to the argument that requiring expert witnesses to confine themselves to their area of responsibility as outlined from the respective legal order is not a deliberate dogmatic choice, but both a logical and procedural necessity. Anything else would amount to trespassing onto the province of the jury.[34]

By arguing that the concept of individualization, salient in forensic science, actually comes down to a decision, and as such hinges on an act of will, it is not denied that a decision can and should be scientifically backed. The point solely is that the scientific model used to articulate the respective target system is only a *conditio sine qua non* for any decision in a system of legal adjudication with commitments to Rationalism. It is not, however, a *conditio per quam* for the respective decision. Expert witnesses inform and educate the fact-finder/juridical decision maker but are not entitled to anticipate their decision. As Lord President Cooper put it, it is the expert witness's duty "to furnish the judge or jury with the necessary scientific criteria for testing the accuracy of their conclusions, so as to enable the judge or jury to form their own independent judgment by the application of these criteria to the facts proved in evidence."[35] This is a matter of actual and contingent and yet valid policy choice, not a misguided and sub-rational pre-scientific operation. Interestingly, even the pioneering forensic scientist Locard supported the view that the laboratory should not become the "antechamber" of the court (Locard, 1940).

Notwithstanding, there is an intrinsic connection between reasoning based on incomplete items of evidence on the one hand, and acts of will on the other hand, leading to decisions: the two instances are connected in the sense that the former is the point of departure of the latter (Biedermann et al., 2008). As much as the scientist cannot interfere with the judicial decision-makers' area of competence, the juridical decision-maker cannot interfere with the process leading to the expert witness testimony, especially its content. The focus, Justice Blackmun remarks, "must be solely on principles and methodology, not on the conclusions that they generate."[36] As long as these decision-structures are a function of valid scientific methods, they cannot *per se* be rejected or disregarded. Scientific conclusions cannot, however, anticipate ultimate issues, i.e., elements of the respective offense to be proved, let alone the verdict as such (Roberts and Zuckerman, 2010, p. 490).

The Role of Formal Theories for Reasoning and Decision Analysis

Through Sections "Transversal Overview of Current Empirical, Normative and Doctrinal Understandings of Expert Witness Testimony" and "Decisional Dimensions of Forensic Expert Testimony" we have aimed at clarifying the intricacies of the systemic relationship between forensic science and the law. Naturally, this raises questions from a variety of viewpoints – normative, descriptive and prescriptive – that, in many discourses on the topic, are not well separated, and hence hinder progress toward a resolution of opposing views.

Many of current debates focus on empirical and descriptive aspects, such as the question of the extent to which witnesses are testifying on the basis of knowledge, and whether the fact-finders can appropriately assess such testimony to reach sound judgments about the disputed events. It is, however, equally important – in our view – to insist on the understanding that structural features of expert witness testimony are actually embedded in a legally structured decision-making process. There are currently two main perspectives in which the decisional dimension of expert witness testimony may be understood.

On an empirical account, claims have been raised that forensic scientists should be subjected to empirical testing. As noted by PCAST, "studies are required, in which many examiners render decisions about many independent tests (typically, involving "questioned" samples and one or more "known" samples) and the error rates are determined" (President's Council of Advisors on Science and Technology [PCAST], 2016, p. 143). Such research leads to general measures of expert performance with respect to a particular area of expertise and/or a given expert's performance. The benefit of this is the provision of information to help assess whether experts, including their methods and techniques, are able to do what they claim to do, and whether they are any better in their tasks than lay persons. The obvious limitation of the empirical perspective is, however, that such general expert performance measures do *not* instruct, normatively, how to make a sound decision (i.e., what to conclude at the end of a forensic examination) in any given individual situation. The latter has to do with the logic of decision and, hence, requires elements of formal methods of reasoning and analysis, among which a prime candidate is (normative) decision theory. Broadly speaking, the purpose of decision theory is to assist decision-makers – in any

[33] The (National Research Council [NAS], 2009, p. 43) remarks on a similar note that the "question is less a matter of whether each person's fingerprints are permanent and unique – uniqueness is commonly assumed – and more a matter whether one can determine with adequate reliability that the finger that left an imperfect impression at a crime scene is the same finger that left an impression . . . in a file of fingerprints."

[34] In view of that distinction Swinton Thomas *LJ* (*R v Davies*) remarked that it is "fundamental that experts must not usurp the functions of the jury in a criminal trial."

[35] *Davie v Edinburgh Magistrates* [1953] SC 34, 40; nota bene this is a Scottish case, which became also the leading authority in England and Wales, see e.g. *R v Gilfoyle* [2001] 2 Cr App R 57, 67, CA.

[36] *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993), Syllabus, at 580.

given case – in thinking about the relative merit of rival courses of action when their outcomes cannot be known with certainty. While this appears to be a good fit for the needs of operational decision makers, there is debate over the possible uses of decision theory since its first presentation in legal literature in the late 1960s. A recurrent critique is that one of the entailments of decision theory, such as expected utility/loss, is not a relevant criterion from a legal point of view because that there are several other aspirations to which the legal process seeks to conform. It is important to note, however, that this is a critique that focuses on the descriptive and prescriptive adequacy of decision theory, while leaving the fundamental question of how to actually make a decision, and justifying it, unresolved. As the eminent figure in decision theory, Howard Raiffa, has noted: "Even if you don't analyze your decision problem by the methodology described in these lectures, you still must act. What will you do?" (Redmayne, 1995, p. 272). We also see here a practical instance of the generality tradeoff mentioned earlier in Section "Law and Philosophy of Science": formal theories assert coherence *at the level of detail* at which they are applied, which depends on the decision analyst's intentions, (computational) capacities and available resources (e.g., in terms of information, time, etc.). It is pointless then to either claim a particular modeling result as a solution for a larger decision problem, or in turn criticize the model for a lack of completeness that it never claimed to have.

We join the above perspective in the sense that it places the inevitability of decision in the first place, and as the overarching perspective (see also Lindley, 1985). This burden of decision, as we call it, has to be absorbed, from which follows the imperative that decision-makers ought to think about their decision problems sensibly, prior to making their decision. The role of formal theories in this task is that of helping individuals make up their minds, in a structured way, about the fundamental ingredients of decision problems (i.e., states of nature, decisions, consequences, etc.). There is nothing prescriptive in this perspective as such, though it provides us with a critical analytical account of current practice. To illustrate this, reconsider one of the currently most controversial forensic reporting issues, that is the problem of *deciding* whether or not a defendant, rather than an unknown person, is the source of, for example, a DNA trace, a partial fingermark or an item of handwriting – a process commonly known as individualization (see Section "Key Controversies Over Selected Aspects of Forensic Expert Testimony"). Our general argument throughout this paper, emphasizing the judiciary's decision-making prerogative, and the imperative to consider all findings, not only scientific findings, is to deny scientists answering this question. Firstly, because it would be an answer provided on an issue (i.e., a proposition of interest), rather than a statement of the value of forensic findings only. Secondly, because deferring the decision to scientists would lead them into an impasse. The impasse is due to the fact that any decision taken in the light of uncertainty is bound to decision consequences, some of which are undesirable (e.g. a false identification), and there is nothing in the scientists' scope of competence that entitles them to assess the relative desirability or undesirability of those consequences (Biedermann et al., 2008, 2016), neither qualitatively and even less so quantitatively. The problematic turn on this is that scientists who continue to make identification decisions, despite this intricacy, will implicitly impose a stance with respect to possible decision consequences to the judiciary, without telling them that they do so, which raises problems of transparency. An even further dimension of concern is that scientists may not even be aware of the decisional dimensions, and their implications, of their form of testimony. Taken together, these intricacies have been recognized as the principal reason why forensic 'identification practice' has become unscientific (Stoney, 1991, 2012).

DISCUSSION AND CONCLUSION

In their paper on the "Individualization Fallacy" Saks and Koehler (2008, p. 215) wonder why so many forensic scientists "ascribe greater powers to their fields than the research supports." A Nietzschean "will to power," masqueraded as individualization claims, lack of understanding[37] for the structure of legal adjudication, the probabilistic (general) character of scientific propositions, or simply an aspiration to 'help solve the crime' are only a few possible answers. Our paper does not aspire to answer this (empirical) question. Further, it is not helpful to fall back in disputes between mainstream evidence scholarship and forensic scientists. Efficient synergy between decision-makers and expert witnesses is too crucial for any modern criminal justice system to be conceptually or even institutionally crippled by a lack of communication and mutual understanding of respective roles and duties of the participants in the legal process. We purported, thus, to clarify – descriptively and analytically – the dimensions in which scientific knowledge, data and related expert assessments manifest themselves in different operative systems, i.e., legal adjudication on the one hand and core scientific theory and practice on the other hand. The conceptual boundary between model-based scientific conjectures and legally contextualized decisions outlines, at the same time, the allocation of epistemic duties and rights between expert witnesses and decision-makers (fact-finders). This perspective diverges from and goes beyond traditional discourses reduced mainly to questions such as admissibility and weight of particular items of scientific evidence because even if the latter issues are settled, the fundamental question on how scientific evidence interfaces with decision making in operational contexts remains an unresolved applied problem. Stated otherwise, even if agreement can be found as to whether an expert witness is appropriately testifying on the basis of knowledge, the fact-finder will still need to intelligently incorporate the witness's testimony in the process of reaching a judgment about the contested events.

When conclusions of forensic scientists do not confine themselves to the scientific findings and their assigned probative value, but amount to categorical assertions about propositions (i.e., 'this person is the source of this crime stain'), and hence represent *local* decisions (to be distinguished from *ultimate* decisions), the precept of factfinders controlling the decision process is violated. As much as ultimate inferential conclusions

[37] Margot (2011, p. 796) has concisely expressed this as follows: "Forensic scientists are proud to see themselves take such an important part in legal proceedings, failing to recognize that they're playing the tune of their masters."

(e.g., about the defendant's liability) are never based solely on the probability of particular propositions of interest, but also involve aspects of juridical classification (Roberts and Zuckerman, 2010, pp. 133–137), conclusions about lower propositional levels (e.g., inference of source; see Section "Key Controversies Over Selected Aspects of Forensic Expert Testimony") involve value judgements regarding the risk of decision consequences (e.g., false identification and, hence, false incrimination of a defendant).

While not prescribing an answer to the above issues, the role of formal methods of reasoning and decision analysis – such as decision theory (Biedermann et al., 2008, 2016) – is to bring these underlying tenets to the open, and to clarify what is fundamentally at stake with any forensic conclusion. The insight from such formal analysis shows, in particular, that the scope and implications of forensic conclusions are much broader than what is commonly thought, because of the required value judgements (e.g., in terms of utilities/losses). The latter call upon a more active role by participants of the legal process other than the scientist. The understanding of forensic expert conclusions in a decisional dimension thus can empower the different parties in the process by showing that the processing of scientific evidence is a task that encompasses broader considerations than those that a scientist alone may address. This is entirely compatible with the view that the ultimate assessment of "any particular piece of evidence, scientific or otherwise, must always be assessed contextually, in the light of its contribution to the case as a whole" (Aitken et al., 2010, p. 70).

It remains the question of what role and position research in judgment and decision making may have in the context the legal adjudication. Over the past decades, literature on this topic has developed extensively and in great depth, and with proponents arguing in controversy about the merits of theoretical research when considering the dynamics of real trials and the limitations of what participants in the legal process are actually capable of doing. These discussions led to valid points to be made from all common analytical viewpoints, normative, descriptive and prescriptive. In this paper, we have extended this perspective to the particular interface between forensic science and the law where, traditionally, the form and content of expert conclusions have attracted critical discussions mainly in a probabilistic perspective, but without giving due consideration to the fact that expert testimony actually amounts to decisions being made with respect to target propositions (e.g., concluding that 'this trace comes from this person'). Rethinking traditional forensic reporting practices, in particular source identifications, in this decisional dimension leads to two main conclusions.

First, while decision-makers in the context of legal adjudication need a scientific basis as a starting point, scientific models and forensic practitioners can at best facilitate the cognitive access to empirical phenomena by providing a systematic account going beyond common knowledge and understanding, i.e., a decision-structure. However, decision-makers need to "jump" (Stoney, 1991, p. 198) in order to render a verdict. As much as model-based propositions (scientific *conclusions*) cannot preempt *decisions* such as the ascription of (criminal) liability, they also cannot preempt decisions regarding forensic source attribution (i.e., concluding that a particular trace or mark comes from a designated person of interest). The main reason for this is that such conclusions depend on more than scientific or other evidence alone. Moreover, modern legal orders choose unequivocally and consistently to allocate the decision-making prerogative to fact-finders (professional judges or jurors), with a clear preference to education over deference.

Second, analyzing expert witness testimony through the lenses of formal theories, in particular normative decision theory, shows that the above allocation of duties and prerogatives actually makes sense; however, the analysis does not claim to practically facilitate the operation of the expert and fact-finder interface. The latter is not a drawback of judgment and decision-making research, but an insight that is valuable to guide ongoing reforms of forensic science reporting practice (expert witness testimony), as evidenced by the recent examples of scholarly works and policy documents drawn from professional bodies and governmental institutions presented throughout this paper.

TABLE OF CASES

Davie v. Edinburgh Magistrates [1953] SC 34, 40.
Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).
Frye v. United States, 293 F. 1013 (D.C. Cir. 1923).
In re Winship, 397 U.S. 358 (1970).
R. v. Deen, The Times January 10.
R v Doheny and Adams [1997] 1 Cr App R 369, 374.
R v Gilfoyle [2001] 2 Cr App R 57, 67, CA.
R v Ireland; R v Burstow [1997] 3 WLR 534.
Wal-Mart Stores, Inc. v. Dukes, et al., 564 U.S. 338 (2011).
Zuckerberg v. Ceglia, U.S. COURT OF APPEALS, 2nd cir., Apr 20, 2015, No. 14-1365-cv.

AUTHOR'S NOTE

This paper has been presented at the 3rd International Symposium on Sino-Swiss Evidence Science ('Pursuing Truth from Different Perspectives') in Hangzhou (China), organized by the China University of Political Science and Law (Beijing, China), Guanghua Law School (Zhejiang University, China) and the School of Criminal Justice (University of Lausanne, Switzerland).

AUTHOR CONTRIBUTIONS

Both authors listed have made equal, substantial, direct and intellectual contributions to the work, and approved it for publication.

FUNDING

This research was supported by the Swiss National Science Foundation through Grant No. BSSGI0_155809 and the University of Adelaide (SA) through a 'Aim for the Stars' Grant 2018.

REFERENCES

Aitken, C., Roberts, P., and Jackson, G. (2010). Fundamentals of Probability and Statistical Evidence in Criminal Proceedings (Practitioner Guide No. 1), Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses, Royal Statistical Society's Working Group on Statistics and the Law, London.

Allen, R. J. (1994). Expertise and the Daubert decision. *J. Crim. L.* & *Criminology* 84, 1157–1175. (1994).

Allen, R. J. (2003). The error of expected loss minimization. *Law Probab. Risk* 2, 1–7. doi: 10.1093/lpr/2.1.1

Allen, R. J. (2015). A note to my philosophical friends about expertise and legal systems. *Hum. Mente J. Philos. Stud.* 28, 71–86.

Bailer-Jones, D. (2009). *Scientific Models in Philosophy of Science*. Pittsburgh, PA: University of Pittsburgh Press. doi: 10.2307/j.ctt5vkdnq

Baron, J. (2008). *Thinking and Deciding*, 4th Edn. New York, NY: Cambridge University Press.

Baron, J. (2012). The point of normative models in judgment and decision making. *Front. Psychol.* 3:577. doi: 10.3389/fpsyg.2012.00577

Biedermann, A., Bozza, S., and Taroni, F. (2008). Decision theoretic properties of forensic identification: underlying logic and argumentative implications. *Foren. Sci. Int.* 177, 120–132. doi: 10.1016/j.forsciint.2007.11.008

Biedermann, A., Bozza, S., and Taroni, F. (2016). The decisionalization of individualization. *Forensic. Sci. Int.* 266, 29–38. doi: 10.1016/j.forsciint.2016.04.029

BKA (Bundeskriminalamt) (2010). *Standard des daktyloskopischen Identitätsnachweises. Vers. 30.06.2010*. Wiesbaden: Bundeskriminalamt.

Buchak, L. (2016). "Decision theory," in *Oxford Handbook of Probability and Philosophy*, eds A. Hájek and C. Hitchcock (Oxford: Oxford University Press), 789–814.

Champod, C., and Evett, I. W. (2001). A probabilistic approach to fingerprint evidence. *J. Foren. Ident.* 51, 101–122.

Champod, C., and Evett, I. W. (2017). *Interpretation, a Personal Odyssey, NIST 2017 Technical Colloquium on Weight of Evidence*. Gaithersburg, MD: NIST, 27–29.

Cole, S. A. (2014). Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the United States. *Law Probab. Risk* 13, 117–150. doi: 10.1093/lpr/mgt014

Cole, S. A. (2018). A discouraging omen: a critical evaluation of the approved uniform language for testimony and reports for the forensic latent print discipline. *Georgia State Univ. Law Rev.* 34, 1103–1128.

Connolly, T. (1987). Decision theory, reasonable doubt, and the utility of erroneous acquittals. *Law Hum. Behav.* 11, 102–112. doi: 10.1007/BF01040444

Criminal Law Revision Committee for England and Wales [CLRC] (1980). *Fourteenth Report, Offences against the Person, Cmnd 7844*. New York, NY: CLRC

Damaska, M. (1973). Evidentiary barriers to conviction and two models of criminal procedure: a comparative study. *Univ. Pennsylvania Law Rev.* 121, 506–589. doi: 10.2307/3311301

Dworkin, R. (1986). *Law's Empire*. Cambridge, MA: Harvard University Press.

Evett, I. W. (2009). Evaluation and professionalism. *Sci. Just.* 49, 159–160. doi: 10.1016/j.scijus.2009.07.001

Evett, I. W., Foreman, L. A., Jackson, G., and Lambert, J. A., (2000). DNA profiling: a discussion of issues relating to the reporting of very small match probabilities. *Crim. Law Rev.* 341–355.

Friedman, R. D. (1997). Answering the Bayesioskeptical challenge. *Int. J. Evid. Proof* 1, 276–291. doi: 10.1111/j.1471-1842.2011.00974.x

Gleick, J. (1998). *Chaos: Making a New Science*. London: Vintage.

Golan, T. (1999). The history of scientific expert testimony in the English courtroom. *Sci. Context* 12, 7–32. doi: 10.1017/S026988970000329X

Haack, S. (2014a). *Federal Philosophy of Science: A Deconstruction–and a Reconstruction, in: Evidence Matters. Science Proof, and Truth in the Law*, Cambridge: Cambridge University Press, 122–155.

Haack, S. (2014b). *Trial and Error: Two Confusions in: Daubert, Evidence Matters. Science Proof, and Truth in the Law*, Cambridge: Cambridge University Press, 104–121.

Hahn, U. (2014). The Bayesian boom: good thing or bad? *Front. Psychol.* 5:765. doi: 10.3389/fpsyg.2014.00765

Hart, H. L. A. (1961). *The Concept of Law*. Oxford: Clarendon Press.

Jackson, J. D. (1988). Two methods of proof in criminal procedure. *Mod. Law Rev.* 51, 549–568. doi: 10.1111/j.1468-2230.1988.tb01772.x

Kaplan, J. (1968). Decision theory and the factfinding process. *Stanford Law Rev.* 20, 1065–1092. doi: 10.2307/1227491

Kaye, D. (1999). Clarifying the burden of persuasion: what Bayesian decision rules do and do not do. *Int. J. Evid. Proof* 3, 1–29. doi: 10.1177/136571279900300101

Kelsen, H. (1934). *Reine Rechtslehre*, 1st Edn. Vienna: Franz Deuticke.

Kirk, P. L. (1963). The ontogeny of criminalistics. *J. Crim. Law Criminol. Police Sci.* 54, 235–238. doi: 10.2307/1141173

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions*, 3rd Edn. Chicago, IL: The University of Chicago Press. doi: 10.7208/chicago/9780226458106.001.0001

Langbein, J. H. (1996). *The Historical Foundations of the Law of Evidence: A View from the Ryder Sources. Faculty Scholarship Series. Paper 551*. Available at: https://digitalcommons.law.yale.edu/fss_papers/551 doi: 10.2307/1123403

Lewis, C. T., and Belanger, C. (2015). The generality of scientific models: a measure theoretic approach. *Synthese* 192, 269–285. doi: 10.1007/s11229-014-0567-2

Lindley, D. (1985). *Making Decisions*, 2nd Edn. Chichester: John Wiley & Sons.

Locard, E. (1940). *L'enquête Criminelle. Traité de criminalistique. Tome septième, Livre VIII*. Lyon: Desvigne.

Lucy, D. (2006). *Introduction to Statistics for Forensic Scientists*. New York, NY: Wiley.

Margot, P. (2011). Commentary on: the need for a research culture in the forensic sciences. *Univ. California Law Rev.* 58, 795–801.

Miller, J. S., and Allen, R. J. (1993). The common law theory of experts: deference or education? *Northwestern Univ. Law Rev.* 87, 1131–1147.

Moretti, T. R., and Budowle, B. (2017). Letter to the editor – Reiteration of the statistical basis of DNA source attribution determinations in view of the Attorney General's directive on "reasonable scientific certainty" statements. *J. Foren. Sci.* 62, 1114–1115. doi: 10.1111/1556-4029.13538

National Research Council [NAS] (2009). *Strengthening Forensic. (Science) in the United States: A Path Forward*. Washington, DC: The National Academies Press.

Oaksford, M. (2014). Normativity, interpretation, and Bayesian models. *Front. Psychol.* 5:332. doi: 10.3389/fpsyg.2014.00332

Office of the Attorney General, U. S. Department of Justice [OAG] (2016). *Memorandum for Heads of Department Components*. Washington, DC: OAG.

Popper, K. R. (1962). *Conjectures and Refutations. The growth of Scientific Knowledge*. New York, NY: Basic Books.

President's Council of Advisors on Science and Technology [PCAST] (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Washington, DC: PCAST.

Raiffa, H. (1968). *Decision Analysis, Introductory Lectures on Choices under Uncertainty*. Reading, MA: Addison-Wesley.

Redmayne, M. (1995). "Doubts and burdens: DNA evidence, probability and the courts," in *Criminal Law Review* 464–482.

Roberts, P., and Zuckerman, A. (2010). *Criminal Evidence*. Oxford: Oxford University Press.

Robertson, B., Vignaux, G. A., and Berger, C. E. H. (2016). *Interpreting Evidence. Evaluating Forensic Science in the Courtroom*, 2nd Edn. Chichester: John Wiley & Sons. doi: 10.1002/9781118492475

Saks, M. J., and Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science* 309, 892–895. doi: 10.1126/science.1111565

Saks, M. J., and Koehler, J. J. (2008). The Individualization Fallacy in Forensic Science Evidence. *Vanderbilt Law Rev.* 61, 199–219.

Sober, E. (2008). *Evidence and Evolution, The Logic Behind the Science*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511806285

Stegmüller, W. (1979). *The Structuralist View of Theories. A Possible Analogue of the Bourbaki Programme in Physical Science*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-95360-6

Stoney, D. A. (1991). What made us ever think we could individualize using statistics? *J. Foren. Sci. Soc.* 31, 197–199. doi: 10.1016/S0015-7368(91)73138-1

Stoney, D. A. (1994). Relaxation of the assumption of relevance and application to one-trace and two-trace problems. *J. Foren. Sci. Soc.* 34, 17–21. doi: 10.1016/S0015-7368(94)72877-2

Stoney, D. A. (2012). Discussion on the paper by Neumann, Evett and Skerrett. *J. R. Statist. Soc.* 175, 399–400.

Suppe, F. (2000). "Understandin scientific theories: an assessment of developments philosophy of science," in *Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association 67, Supplement, Part II: Symposia Papers* (Chicago, IL: University of Chicago Press), S102–S115.

Taroni, F., Biedermann, A., and Bozza, S. (2016). Statistical hypothesis testing and common misinterpretations: should we abandon p-value in forensic science applications? *Forensic Science International* 259, e32–e36. doi: 10.1016/j.forsciint.2015.11.013

Tarski, A. (1944). The semantic conception of truth: and the foundations of semantics. *Phil. Phenomenol. Res.* 4, 341–376. doi: 10.2307/2102968

Thayer, J. B. (1892). *Select Cases on Evidence at the Common Law*. Cambridge, MA: C.W. Sever.

The Council of the Inns of Court [COIC] and The Royal Statistical Society [RSS] (2017). *Statistics and Probability for Advocates: Understanding the use of Statistical Evidence in Courts and Tribunals*. London: RSS.

Thompson, W. C., and Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defense attorney's fallacy. *Law Hum. Behav.* 11, 167–187. doi: 10.1007/BF01044641

Toulmin, S. (1967). *Philosophy of Science. An Introduction*. London: Hutchinson.

Twining, W. (1982). "The rationalist tradition of legal scholarship," *Well and Truly Tried: Essays on Evidence in Honour of Sir Richard Eggleston*, eds E. Campbell and L. Waller (Sydney, SA: Law Books), 211–249.

U. S. Department of Justice [DOJ] (2018). *Approved Uniform Language for Testimony and Reports for the Forensic Latent Discipline (Latent Print ULTR).* Washington, DC: DOJ.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Oxford University Press.

Wigmore, J. H. (1908). *A General Survey of the History of the Rules of Evidence, in: Select Essays in Anglo-American Legal History,* Vol. 2. Boston, MA: Little, Brown & Co.

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: 07 January 2019
doi: 10.3389/fpsyg.2018.02634

Better Together: Reliable Application of the Post-9/11 and Post-Iraq US Intelligence Tradecraft Standards Requires Collective Analysis

Alexandru Marcoci[1,2], Mark Burgman[3], Ariel Kruger[4], Elizabeth Silver[5], Marissa McBride[3,4], Felix Singleton Thorn[5,6], Hannah Fraser[4], Bonnie C. Wintle[4,7], Fiona Fidler[4,5] and Ans Vercammen[3,4]*

[1] Department of Philosophy, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [2] Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, London, United Kingdom, [3] Centre for Environmental Policy, Imperial College London, London, United Kingdom, [4] School of Biosciences, University of Melbourne, Parkville, VIC, Australia, [5] School of Historical and Philosophical Studies, University of Melbourne, Parkville, VIC, Australia, [6] Melbourne School of Psychological Sciences, University of Melbourne, Parkville, VIC, Australia, [7] Centre for the Study of Existential Risk, University of Cambridge, Cambridge, United Kingdom

OPEN ACCESS

Edited by:
Nathan Dieckmann,
Oregon Health & Science University,
United States

Reviewed by:
James Kajdasz,
United States Air Force Academy,
United States
Daniel Benjamin,
University of Southern California,
United States

***Correspondence:**
Alexandru Marcoci
marcoci@unc.edu;
a.marcoci@lse.ac.uk

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

Received: 31 August 2018
Accepted: 07 December 2018
Published: 07 January 2019

Citation:
Marcoci A, Burgman M, Kruger A, Silver E, McBride M, Singleton Thorn F, Fraser H, Wintle BC, Fidler F and Vercammen A (2019) Better Together: Reliable Application of the Post-9/11 and Post-Iraq US Intelligence Tradecraft Standards Requires Collective Analysis. Front. Psychol. 9:2634. doi: 10.3389/fpsyg.2018.02634

Background: The events of 9/11 and the October 2002 National Intelligence Estimate on Iraq's Continuing Programs for Weapons of Mass Destruction precipitated fundamental changes within the United States Intelligence Community. As part of the reform, analytic tradecraft standards were revised and codified into a policy document – Intelligence Community Directive (ICD) 203 – and an analytic ombudsman was appointed in the newly created Office for the Director of National Intelligence to ensure compliance across the intelligence community. In this paper we investigate the untested assumption that the ICD203 criteria can facilitate reliable evaluations of analytic products.

Methods: Fifteen independent raters used a rubric based on the ICD203 criteria to assess the quality of reasoning of 64 analytical reports generated in response to hypothetical intelligence problems. We calculated the intra-class correlation coefficients for single and group-aggregated assessments.

Results: Despite general training and rater calibration, the reliability of individual assessments was poor. However, aggregate ratings showed good to excellent reliability.

Conclusion: Given that real problems will be more difficult and complex than our hypothetical case studies, we advise that groups of at least three raters are required to obtain reliable quality control procedures for intelligence products. Our study sets limits on assessment reliability and provides a basis for further evaluation of the predictive validity of intelligence reports generated in compliance with the tradecraft standards.

Keywords: intelligence analysis, intelligence failures, intelligence reform, IRTPA, ICD203, ODNI, tradecraft standards, inter-rater reliability

INTRODUCTION

In a seminal article on the role of intelligence analysis, Betts wrote that "the role of intelligence is to extract certainty from uncertainty and to facilitate coherent decision in an incoherent environment" (Betts, 1978, p. 69). In other words, the role of intelligence analysis is to *reduce* (but not necessarily eliminate, see Heazle, 2010; Marrin, 2012) and *caveat* uncertainty (Friedman and Zeckhauser, 2012) to improve national security policy. However, in the wake of perceived intelligence failures such as predicting the 9/11 attacks (National Commission on Terrorist Attacks Upon the United States, 2004, but also see Zegart, 2005, 2006) and Iraq's capability of deploying weapons of mass destruction (The Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction, 2005, but also see Phythian, 2006), the intelligence community's (IC) ability to help policy makers manage uncertainty was criticized. In consequence, the United States Congress passed sweeping reforms in the 2004 Intelligence Reform and Terrorism Prevention Act (IRTPA), demanding, among other things, the adoption of analytic tradecraft standards to improve the quality of reasoning and argumentation in intelligence products. Moreover IRTPA mandated the creation of an ombudsman for analytic integrity to ensure "finished intelligence products produced [. . .] are timely, objective, independent of political considerations, based upon all sources of available intelligence, and employ the standards of *proper analytic tradecraft*" (IRTPA, 2004, Section 1019a, our emphasis).

In response, the Director of National Intelligence signed Intelligence Community Directive [ICD] 203 (2007/2015), specifying four analytic standards: objectivity, political independence, timeliness, and good tradecraft. The latter further identifies nine elements of analytic tradecraft: (1) Properly describes quality and credibility of underlying sources, data, and methodologies; (2) Properly expresses and explains uncertainties associated with major analytic judgments; (3) Properly distinguishes between underlying intelligence information and analysts' assumptions and judgments; (4) Incorporates analysis of alternatives; (5) Demonstrates customer relevance and addresses implications; (6) Uses clear and logical argumentation; (7) Explains change to or consistency of analytic judgments; (8) Makes accurate judgments and assessments; and (9) Incorporates effective visual information where appropriate. To ensure compliance with these standards, an office for Analytic Integrity and Standards (AIS) was established in the Office of the Director of National Intelligence.

Nevertheless, the belief that compliance with these standards would improve analysis and reduce uncertainty has been challenged from two directions. First, it has been claimed the tradecraft standards in ICD203 add nothing new to the existing practice (Lowenthal, 2012; Marchio, 2014; Gentry, 2015), and second, that reports complying with these standards may not produce more accurate estimates (Tetlock and Mellers, 2011). In this paper we investigate a third, more fundamental, issue that has so far received very little attention: whether the tradecraft standards can be reliably applied; that is, will two (or more) assessors evaluating the same report reach the same conclusions regarding its quality? (see Marcoci et al., 2018).

There is reason to be concerned. First, research into the design and implementation of assessment standards and requirements in higher education show consistently that standards expressed in linguistic terms are "fuzzy" and subject to multiple interpretations even by experienced evaluators (Sadler, 1987; Freeman and Lewis, 1998; Webster et al., 2000; O'donovan et al., 2004). ICD203 shares many of the characteristics of assessment standards in higher education.

Second, individual expert judgments in many related fields are routinely insufficiently reliable for practical applications. The reliability of judgments about facts and future events correlates poorly or not at all with the personal attributes that conventionally are associated with an expert's credibility such as qualifications, years of experience, memberships, publications, or the esteem in which they are held by their peers. Cooke (1991) was one of the first to explore the ramifications of expert uncertainty for safety systems in engineering. Since that seminal work, hundreds of publications in spheres ranging from medicine and ecology to safety engineering and geoscience have documented the difficulties of identifying the attributes of reliable raters and the benefits of using group judgments to improve reliability (Burgman, 2015).

We could not find any evidence of attempts to identify the best assessors of analytic products or any research into the impact of training on their performance. To our knowledge, the reliability of the analytic tradecraft standards has not been systematically assessed. Yet, the question of reliability logically precedes investigation about its construct validity (do the standards really capture good quality of reasoning?), and predictive validity (does a good report make accurate predictions about the state of the world?). If the standards are not construed and used in the same way by different users, then the question of whether they engender more accurate estimates becomes moot. In this paper we report on the results of an experiment gauging the reliability with which the tradecraft standards in ICD203 can be applied. As noted above, ICD203 is meant to direct both the production and evaluation of analytic reports for quality control. For purposes of this experiment we focus on the latter aspect of ICD203.

MATERIALS AND METHODS

Participants

We recruited 15 participants through an advertisement posted on the University of Melbourne's School of Historical and Philosophical Studies mailing list. Selection criteria included: (1) completed or currently enrolled in a research higher degree in Arts/Humanities, (2) experience marking essays, (3) interest in the study and availability/willingness to work under imposed time constraints. Participants were selected from a pool of applicants based on best fit to the selection criteria, and were remunerated for their time. Seven were male, seven female and one preferred not to specify. Their average age was 35.93 (SD = 9.14) years. Seven had completed either a Masters or a PhD in the Humanities, while the rest were current PhD candidates.

Materials and Procedures

This study draws on materials developed in the *Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE)* program, an active research program (2017–2020) run by the Intelligence Advanced Research Projects Activity (IARPA).

Reports

Participants (henceforth: "raters") were asked to evaluate the quality of reasoning of a set of hypothetical intelligence reports. The reports were generated by another group of research participants involved in testing a new online collaborative reasoning platform developed at the University of Melbourne, Australia, as part of the CREATE program. This platform (described in van Gelder and de Rozario, 2017), aims to (a) use the power of distributed processing within a network of individual thinkers, and (b) improve reasoning quality and the aggregation of solutions into a final, agreed solution. Users on the platform are requested to write individual analytical reports that outline the outcome (the solution to the problem) and the process (the underlying reasoning). They are invited to comment on one another's contributions and update their own contribution in response to comments. The platform also encourages users to rate others' solutions and the average quality rating (on a scale 0–100) determines the rank of each solution. The top-rated solution becomes the template for a final draft report, which is edited and ultimately submitted as a team report.

We collected 64 such reports generated by both individual users and teams in response to four different reasoning problems. All problems emulated reasoning challenges in real intelligence problems, with the exception that the problems were self-contained, i.e., all necessary information for solving them was contained in the problem description and the contextual information provided. Reports were generated during "beta-testing" of the platform in late 2017; test users included platform developer team members, junior analysts from an intelligence organization and individuals recruited online via targeted Facebook advertising. For the purposes of the current study, reports were downloaded from the platform and formats retained apart from minor changes such as removing specific references to and comments on other users' submissions, so that each report could be analyzed as a stand-alone item. An example problem and report are provided in **Supplementary Material**.

The number of reports included in this study was determined prior to data collection. Given our knowledge of the internal procedure in AIS, we decided to measure the inter-rater (rather than intra-rater) reliability of the tradecraft standards using intra-class correlations (ICC). As the precision of ICC estimates depends on the number of raters, the number of reports and the true ICC, the numbers of included raters and reports were determined *a priori* to ensure sufficient precision (i.e., narrow confidence intervals) in our estimates of the ICC following Bonett (2002). With 64 products and 15 raters we were certain to have a 95% CI width (distance between the upper and lower bound) of no larger than 0.19 for the average ICC, and only slightly wider intervals in our estimates of the ICC for fewer raters (e.g., at worst 0.25 for estimating the ICC with four raters), regardless of the true ICC (Bonett, 2002).

Quality of Reasoning Rubric

ICD203 outlines the standards for good reasoning in intelligence analysis. These standards are operationalized by AIS in a "Rating Scale for Evaluating Analytic Tradecraft Standards," an assessment rubric with nine criteria (**Table 1**). The rubric is very detailed. Every criterion has a short explanation regarding its scope. For example, "Criterion 4 – Incorporates analysis of alternatives" gives a paragraph of explanation detailing what it takes for a report to "incorporate analysis of alternatives." Further, every criterion includes a comprehensive description of four levels of performance quality, i.e., "poor," "fair," "good," and "excellent," except for criterion 7, which describes the quality of judgments present in the report as either "unclear," "conditioned," or "unconditioned." Each level of performance contains detailed sub-criteria that a report must meet to count as having satisfied the criterion "up to that level." Additionally, each criterion contains a "Notes" section that gives detailed examples, tips, hints and elements to "watch out for" when applying the rubric. These three elements (high level explanation, sub-criteria for each level of satisfaction, and the notes) sum to a detailed rubric.

For the purposes of this study, we omitted the criterion "Explains change to or consistency of analytical judgments" because it requires the report writer to have an understanding of previous analyses, which is irrelevant for the kind of constrained reasoning problems we used in this study. Furthermore, we made a small number of minor textual changes to accommodate the participants' lack of familiarity with the jargon used in the original document. Numeric values were assigned to each performance level (i.e., 0 for "poor," 1 for "fair," 2 for "good," and 3 for "excellent," except for criterion 7, where 0 was awarded for "unclear," 1 for "conditioned" and 2 for "unconditioned"). Scores were summed to give a total mark out of 31 for each report.

Procedures

All raters completed the rating of the 64 reports over the course of 4 working days, in supervised sessions held at Melbourne University. This allowed us to mitigate the risk of non-independent evaluation, manage rater fatigue, ensure that raters understood the instructions, and that each report was

TABLE 1 | Criteria used for assessing quality of reasoning in the rating scale for evaluating analytic tradecraft standards.

Criterion	Description
1	Properly describes quality and credibility of underlying sources, data, and methodologies
2	Properly expresses and explains uncertainties associated with major analytic judgments
3	Properly distinguishes between underlying intelligence information and authors' assumptions and judgments
4	Incorporates analysis of alternatives
5	Demonstrates relevance and addresses implications
6	Uses clear and logical argumentation
7	Makes accurate judgments and assessments
8	Incorporates effective visual information where appropriate

marked in full. Compliance with instructions was monitored by one of the authors (AK).

The rating process started with a 2-h training/calibration exercise led by one of the authors (AK). First, raters were given the Rating Scale to peruse, make notes and ask questions regarding anything that was unclear or ambiguous. Second, raters were given a copy of a sample hypothetical intelligence problem (not included in the experiment, but also drawn from the CREATE program) and again encouraged to peruse it and ask questions. Next, raters were split into five groups of three and asked to evaluate a sample report individually at first and then deliberate in their group to reach a consensus on its evaluation using the Rating Scale. Afterward, groups shared their evaluations, followed by a robust discussion that highlighted differences in the way each group had interpreted the rubric criteria. Through facilitated discussion, differences were resolved and raters reached a consensus on an interpretation of how the criteria should be applied. They repeated this process for another two sample reports on the same hypothetical intelligence problem. Finally, the group was given one last sample report to mark as individuals and the facilitator assessed whether the group was sufficiently calibrated. At the end of this process, raters appeared to have reached a shared praxis or understanding of how to apply the rubric to the types of reports they would be evaluating. Each rater was then presented with a bound book containing the 64 reports in randomized order to eliminate order effects. Raters indicated their assessments on a personal score sheet, and were instructed not to discuss the reports, the rubric or the ratings with each other (data collected is summarized in the **Supplementary Material**).

Upon completion of the 64 report ratings, we obtained feedback from the raters on their experience with the rubric, and how they thought it performed as an assessment tool of quality of reasoning. All participants completed a questionnaire consisting of 10 open-ended and 10 multiple-choice questions, and took part in a focus group session (3 h). Both the survey and the focus group explored what the raters thought did and did not work well, which criteria were difficult to apply and why, whether there were elements of good or bad reasoning that were not captured by the rubric, the user-friendliness of the rubric, and their confidence in applying the rubric for the assessment of reasoning quality (see **Supplementary Material**, **Table 3** for the full list of questions).

Analysis

Rubric Reliability

The AIS quality control procedure involves multiple assessors concurrently evaluating products on the basis of ICD203. This motivated the use of the inter-rater reliability as our primary metric for rating consistency. Inter-rater reliability of the summed total scores for each report was assessed via ICC, a commonly used metric for the reproducibility or "consistency" of quantitative measurements made by different observers rating the same object(s). ICC values lie between 0.0 and 1.0, with higher values corresponding to greater agreement between raters.

First, we used the IRR package in R to calculate ICC values using a Two-Way Random-Effects Model, which assumes that each object is rated by a different set of raters who were randomly chosen from a larger population of possible raters. The ICC value we report here reflects absolute agreement rather than simple consistency between raters. We report both the "average" ICC value and the "single" ICC value, which differ in their interpretation. Their use depends on how the measurement protocol will be conducted in actual application. The "single" ICC is an index for the reliability of the ratings of single raters; the "average" ICC is an index for the reliability of different raters averaged together. The latter always results in higher ICC estimates. If in future use of the rubric, the average value across a number of raters is used as the assessment basis, the relevant reliability metric would be the "average" ICC. Conversely, if in future applications of the rubric, a single rater conducts the actual assessment, the "single" ICC type is the relevant reportable metric, even though the reliability study involves two or more raters. Regardless of the type of ICC, values <0.40 indicate poor inter-rater agreement, between 0.40 and 0.59 fair agreement, between 0.60 and 0.74 good agreement and >0.75 excellent agreement (Cicchetti, 1994).

Second, we examined the internal consistency of the eight criteria that make up the rubric with Cronbach's Alpha. We also assessed item-total correlations to examine which (if any) criteria showed poor consistency with the rest of the rubric. Criteria with poor item-total correlations should be considered for removal from the rubric as they compromise reliability.

Rater Feedback

Results from the rater survey were summarized with descriptive statistics by one of the authors (BW). With regard to the focus group, two of the authors (AK, MM) independently coded the transcript and extracted the main themes. The resulting themes were reviewed by four authors (AK, AM, AV, and MM) to ensure that each theme was internally coherent, themes were distinct, and to reach consensus on their naming and interpretation.

RESULTS

Inter-Rater Reliability

We calculated the ICC value for groups of raters of varying size. We first examined the "average ICC" metric for groups of between 2 and 15 raters. The "average ICC" provides a valuable estimate of reliability if future applications of the rubric involve aggregated evaluations, that is, if multiple raters are tasked with assessing single reports and their scores are averaged to produce a final quality assessment.

To ensure reliability of our findings, we iterated over all possible subsets of each given group size n, and report the average ICC values for each n. We found an increase in reliability with increasing numbers of raters, starting from fair reliability with $n = 2$ raters [ICC = 0.498, 95% bootstrap CI = (0.196, 0.799)] to close to perfect reliability [ICC = 0.897, 95% CI = (0.846, 0.936)] when $n = 15$ raters were included (**Figure 1**). However, even a small set of three raters produces borderline good reliability [ICC = 0.608, 95% bootstrap CI = (0.416, 0.800)]. On the other hand, the "single ICC" metric produces an estimate of the

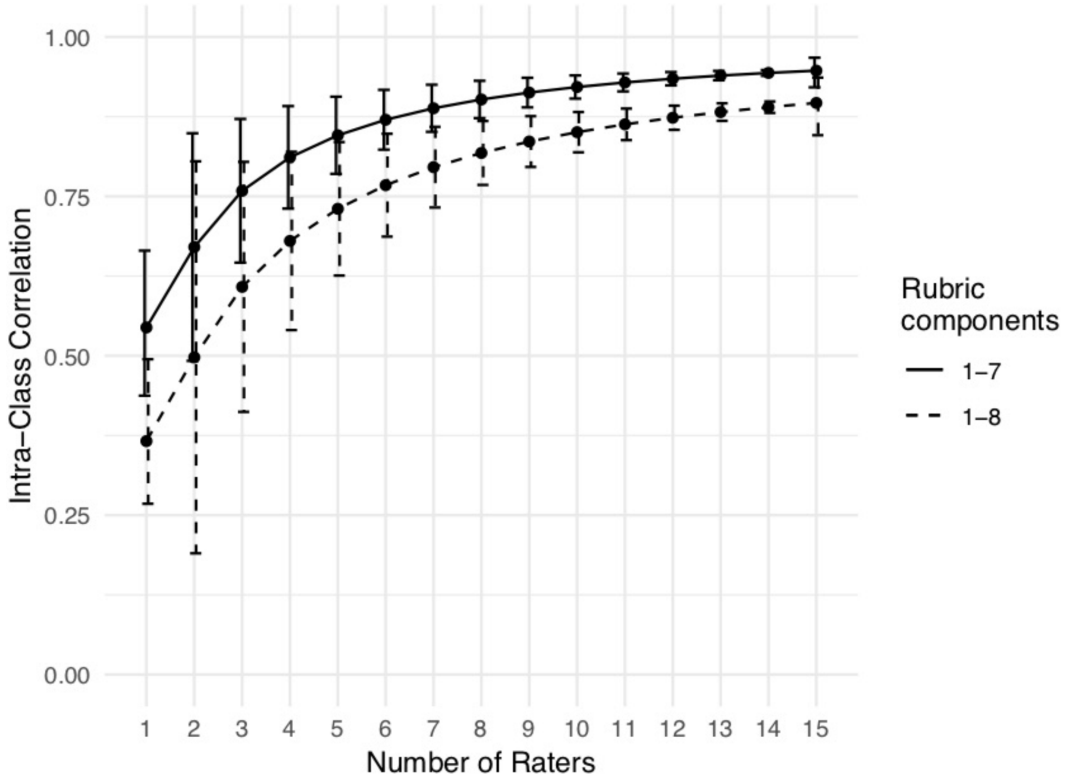


FIGURE 1 | Relationship between number of raters and the Intra-Class Correlation (ICC), using either components 1–8 of the rubric, or only components 1–7. For $n = 1$ rater, the ICC estimate is based on an estimate of "single ICC" as calculated by the IRR package, using data from all 15 raters. For $n = 2$ to $n = 15$, to obtain the most precise estimates, we calculated the "average ICC" for all subsets of raters of size n, and took the mean. Error bars are 95% confidence intervals. For $n = 1$ and $n = 15$, the 95% confidence interval was calculated by the IRR package when estimating the ICC value. For $n = 2$ to $n = 14$, the confidence intervals are 95% bootstrap intervals over the subsets of raters.

reliability of a scale if just one rater scored it on a single occasion. Accordingly, this value does not depend on number of raters in the group, and we find that based on the 15 available raters, the single ICC value was poor [ICC = 0.366, 95% CI = (0.268, 0.495)].

Removing criterion 8 ("Incorporates effective visual information where appropriate") from the overall score calculation improved the ICC values almost as much as doubling the number of raters (For example, using the complete rubric, three raters have an ICC of 0.608. Using six raters would improve the ICC to 0.768, whereas keeping only three raters but dropping criterion 8 improves the ICC to 0.759).

We also assessed the inter-rater reliability for individual criteria. The average ICC for 15 raters was excellent for all but criterion 8. The single ICC value for 15 raters, however, was poor to fair across the criteria (**Figure 2**). For a detailed analysis of the ICC values for all possible group sizes, see **Supplementary Material** (**Figure 3**).

Internal Consistency

We calculated Cronbach's Alpha separately for each of the 64 reports. The average value of Alpha across all reports was 0.606, which is considered "questionable" internal consistency. To examine which of the criteria might be responsible for the low consistency, we conducted an item-total correlation analysis (**Table 2**). This revealed that criteria 7 ("Makes accurate judgments and assessments") and use 8 showed little correlation with the remaining items, and that removing them would improve the internal consistency of the rubric, with Alpha = 0.64 and 0.66, respectively. Rerunning the internal consistency analysis with both items seven and eight removed, revealed that this increased the internal consistency across all reports to "acceptable," Alpha = 0.71.

The removal of criterion 8 in particular appears to be defensible as the inter-rater reliability for this criterion was very low, and it was identified in our qualitative analysis as neither critical to "good reasoning" nor transparent in its application, at least with reference to the specific set of test reasoning problems in this study (refer to **Supplementary Material**, **Figure 4** for further details).

Qualitative Feedback

The results from the qualitative feedback reveal that some criteria are ambiguous and provide insufficient guidance, allowing for potential discrepancies in interpretation. Moreover, some criteria lack specificity, that is, raters perceived areas of overlap such that judgments on some criteria depended on and were affected by

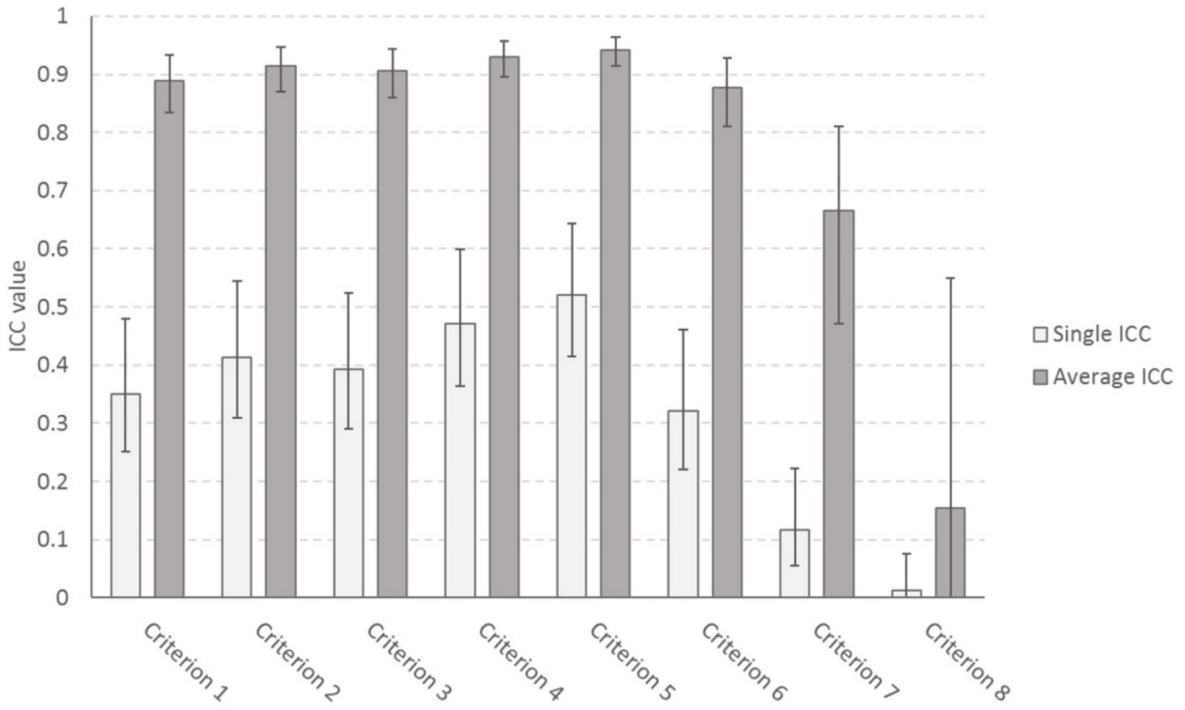


FIGURE 2 | Intra-class correlation coefficients for each of the criteria of the assessment rubric based, with 95% confidence intervals.

others. Some criteria, in particular 1 ("Properly describes quality and credibility of underlying sources, data, and methodologies") and 2 ("Properly expresses and explains uncertainties associated with major analytic judgments"), were also considered to describe multiple distinct attributes, potentially leading to conflation of reasoning faults in the overall assessment.

The following themes emerged as the key points of concern regarding the rubric used in this experiment. Illustrative verbatim quotes from the transcript are included for each theme:

1. "Box-ticking." In applying the rubric, marks are assigned for the presence of certain attributes rather than on the actual quality or effectiveness of these elements.

"But if Y provides what we would think would be very weak evidence for the truth of X then for a certain kind of box ticker that arguably would be enough to qualify it as good but for someone who's perhaps more quality minded and perhaps is arguably inclined to go beyond the rubric they might say - no that counts as poor."

"And I felt it rewarded just putting headings and separating information. It gave too much value to just distinguishing when it was done very blatantly and not very well."

2. "Granularity." Descriptions for the different levels of satisfaction for a given criterion were not precise enough to enable clear categorization between poor-fair-good-excellent.

"I often had trouble distinguishing between fair and good . . . I wanted a third option in between cause they might provide say 1 sentence that's obviously little detail but if they go into 3 or 4 sentences I wouldn't call it considerable but you have to go one or the other."

3. "Specificity." Some criteria were too dense and measured multiple attributes at once (that may diverge in a single report). This may also have led to perceived overlap between criteria.

". . . I found that reports that did perform in the excellent category in criterion 3 usually automatically perform well in criterion 4 as well because the two things kind of go together."

4. "Logical consistency." A number of comments identified logical inconsistencies, or a lack of overall coherence in the criteria in how they addressed the overall goal of "good reasoning."

TABLE 2 | Item-total correlations for the eight criteria, and estimated Alpha values if the criterion were removed from the rubric.

Criterion	Item-total correlation	Cronbach's Alpha, if deleted
Criterion 1	0.41	0.58
Criterion 2	0.46	0.57
Criterion 3	0.51	0.56
Criterion 4	0.38	0.59
Criterion 5	0.40	0.57
Criterion 6	0.53	0.55
Criterion 7	0.19	0.64
Criterion 8	0.09	0.66

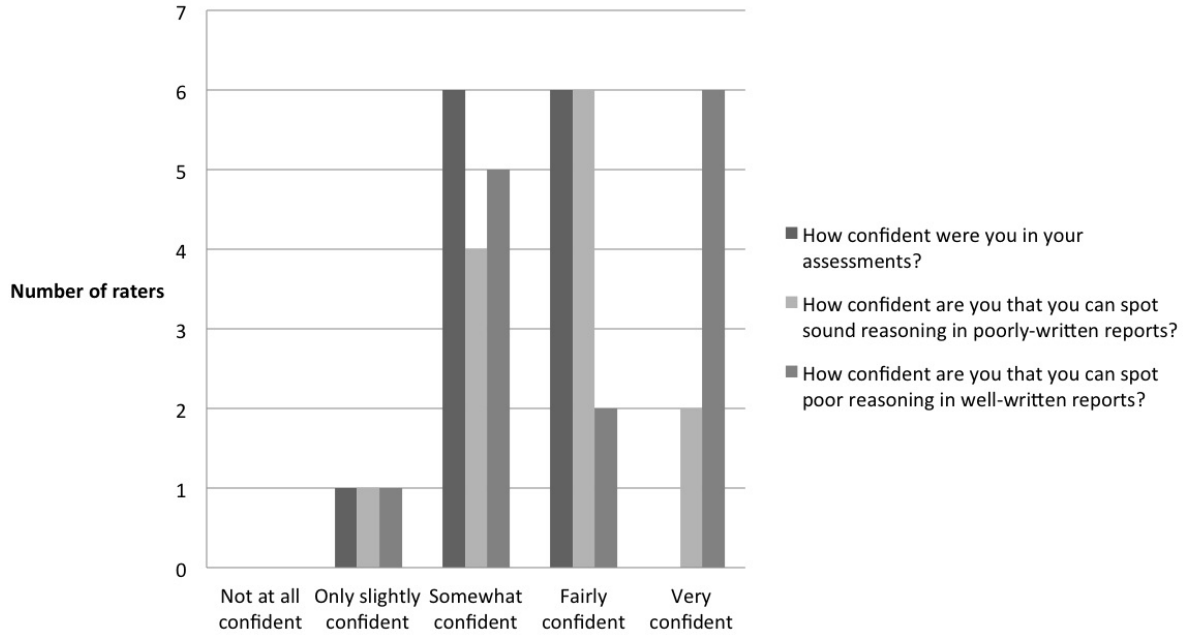


FIGURE 3 | Raters' evaluation of overall confidence in their report assessments (n = 13).

"…still scored really highly even though . . . wasn't addressing the question at all, because it presented a clear analytic message,…did it really well, but it should be a problem if it's not answering the actual question."

"… it's a little bit inconsistent . . . in the notes it says you're not to rate reports on writing style or editorial practices. Seems a bit contradictory. But also I do lot of work as an editor and I would argue that writing style and editorial practices are connected to clarity and logic of argument. The way you use words, punctuation could change the meaning of a sentence."

5. "Aggregation." The rubric provides no detail on how criterion scores are aggregated into a total score. Perceived weighting of the different criteria influenced rater behavior that may introduce inconsistency in the application of the rubric.

"If you're going to use rubrics you need to understand how each criterion is weighted because some are literally more important than others . . .. It kind of depends on what the actual goal is as to how things are weighted."

6. "Unfair penalization." Raters felt forced to penalize a report for "going the extra mile," rather than rewarding some risk taking.

"It's confusing that if you say you make a claim about probability and don't explain it all you get fair whereas if you explain it wrongly…you get penalized."

Overall, raters reported being moderately confident in their report assessments, with the majority being "fairly" or "somewhat" confident (**Figure 3**).

When asked how useful (or not) the rubric is in helping to separate quality of reasoning with quality of writing (n = 14), eight found it to be "fairly useful" (57%), four found it to be "somewhat useful" (29%), and two found it "only slightly useful" (14%).

DISCUSSION

Our study clearly illustrates that the tradecraft standards and their operationalization in the AIS rubric present "poor" inter-rater reliability when deployed by individual raters and "good" to "excellent" (when criterion 8 is excluded) inter-rater reliability when deployed by groups of at least three raters. We expect this will be true of reports of the kind used in this study: i.e., those that are relatively simple and self-contained. In contrast, a group of 15 raters approaches perfect reliability. Therefore, our results suggest that evaluations completed by a single assessor on the basis of the tradecraft standards should be interpreted with extreme caution. This study focused exclusively on the role of the tradecraft standards in the *evaluation* of analytic reports. But the low inter-rater reliability of the tradecraft standards when used by single raters also raises in our minds concerns regarding their use in the *production* of analytic reports.

Our findings further indicate that the criteria are not sufficiently precise, are ambiguous, may not be exhaustive in capturing the core elements of good reasoning, and may be perceived by analysts as not applicable in the development of a well-reasoned report. Users may also perceive the standards to be emphasizing "process" over "deep quality." The criteria should therefore be revised to ensure that they are internally consistent and that each addresses a single issue. Moreover, in the absence

of knowledge on how the criteria scores will be aggregated, raters may hold private and different weightings that influence their assessments of the products. Similar concerns may arise in the minds of the analysts who *produce* reports using ICD203. This may lead to a focus on report attributes that are of limited relevance to the overall quality or the accuracy of the analysis.

Further research is required to provide a comprehensive evaluation of the success of IRTPA and ODNI/AIS in creating a reliable and valid quality control process for the IC. This study is a first step in that direction and has its limitations.

First, note that a group's rating of a report consists in the mathematical average of the individual ratings. This approach cancels out disagreements. Another approach would be to require discussion and/or third party moderating to resolve disagreements on the application of the standards before individual ratings are averaged. Whether such an approach would raise the reliability of teams of evaluators (and by how much) remains an open empirical question that we aim to address in future research. And the potential for bias should not be overlooked. Nevertheless, the fact that small (n = 3–4) teams can apply the standards consistently (even when using simple mathematical averaging) means teams of evaluators using the current AIS operationalization of ICD203 can perform reliable quality control.

Second, our raters were novices in the sense that they had no prior experience in using either the tradecraft standards or the AIS rubric. However, they had considerable experience using assessment rubrics in higher education and assessing written work, they were given training on the AIS rubric and underwent a calibration exercise. Given the strong parallels between rubrics in education and this one, there are no obvious reasons to expect that novice professional raters would perform appreciably better. Moreover, the literature on expertise teaches us that the attributes of reliable evaluators are very elusive (Burgman, 2015). So whether the results of the present study would hold for senior assessors on real intelligence reports remains an open (empirical) question. On the one hand, due to experience, they may be more consistent in the application of the standards. On the other, "real-world" (unconstrained) intelligence problems and reports would be more difficult to assess, and the impact of idiosyncratic understanding of the standards of analysis and biases should not be underestimated. It is unclear to us how to weigh these considerations *a priori* and we hope to address the reliability of the standards with senior assessors and on unconstrained intelligence problems in future research.

Third, we should also note that, all other factors being equal, ICC values are depressed when there is little variation in the objects being rated. If reports used in this study were relatively similar in quality, this would therefore impose an upper limit to the achievable ICC. However, the reports represented a range of products generated on the platform, by both individuals and teams, and vary in sophistication and reasoning quality. This was confirmed by examining the peer assessments produced by contributors themselves using the optional rating functionality on the platform. While only a subset of reports (N = 40) received peer-ratings, the quality assessments ranged from 10 to 85 on a 100-point scale (M = 57.4, SD = 14.45), suggesting sufficient variation in quality for the purposes of ICC calculation. Future evaluation of the tradecraft standards should nevertheless be performed on a wide range of reports varying in style, purpose and quality.

Furthermore, whether *accurate* quality control is possible on these standards remains an important open question. Just because averages of groups of three or more raters are consistent does not mean that their assessments accurately capture the true quality of reasoning. This is a matter of external validity. Validity is dependent on reliability – an unreliable instrument cannot make accurate measurements – hence, this study should be considered a first step toward an investigation of the validity of the tradecraft standards. But the matter of whether these standards and the associated rubric used by AIS is actually a valid indicator of quality of reasoning, and whether a report that rates highly is also producing the "correct" results is one for future study.

Finally, in-depth analysis of the 15 raters' experience in applying the rubric revealed potential leverage points to revise the instrument with a view to increasing its internal consistency. Some criteria were too prescriptive, described as a "box-ticking exercise," leading to frustration. Raters felt that if they complied with the rubric, they were forced to unfairly penalize genuine, though incomplete, analytical process, whereas the absence of analytic effort was rewarded, comparatively. In a context where analysts may already feel under pressure to align with a preferred narrative (Stimson and Habeck, 2016), this may promote a culture of conservative analytical approaches at the expense of appropriate risk-taking, which may be detrimental to the overall quality of reports.

Commenting on the intelligence reform brought about by IRTPA, Robert Cardillo (2010), who served as the Deputy Director of National Intelligence for Intelligence Integration, wrote that ICD203 "injects rigor into our processes and products and holds analysts and managers accountable for results" (2010: 44), i.e., by providing a tool for assessing the analytic products they generate. The results of the present study suggest that this optimism may be compromised when evaluations are undertaken by single assessors, but that it may be vindicated by teams who can consistently apply the tradecraft standards to evaluate the quality of products generated by the IC.

ETHICS STATEMENT

This research project has been approved by the Human Research Ethics Committee of The University of Melbourne, with ethics ID number 1646872.5. All participants in this study were employed as casual research assistants and signed a work contract. Participants who produced the reports used in this study have signed an informed consent form.

AUTHOR CONTRIBUTIONS

AM and AV designed the experiments with the contributions of MB, FST, HF, BW, and FF. AK conducted the experiments.

AK and MM coded the transcripts from the focus group, and BW analyzed the survey data. AV and ES analyzed the quantitative data. AM and AV took the lead in writing the manuscript. All authors provided critical feedback and helped to shape the research, analysis and manuscript.

FUNDING

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) (2016), under Contract (16122000002). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the United States Government. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

ACKNOWLEDGMENTS

We would like to thank Tim van Gelder, and our many colleagues in the Melbourne CREATE research team for their comments, suggestions and advice. We would also like to thank our two reviewers for comments that improved the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02634/full#supplementary-material

REFERENCES

Betts, R. (1978). Analysis, war, and decision: Why intelligence failures are inevitable. *World Polit.* 31, 61–89. doi: 10.2307/2009967

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Stat. Med.* 21, 1331–1335. doi: 10.1002/sim.1108

Burgman, M. A. (2015). *Trusting Judgements: How to Get the Best out of Experts*. Cambridge: Cambridge University Press.

Cardillo, R. (2010). A cultural evolution. *Stud. Intell.* 54, 43–49.

Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instrument in psychology. *Psychol. Assess.* 6, 284–290. doi: 10.1037/1040-3590.6.4.284

Cooke, R. M. (1991). *Environmental Ethics and Science Policy Series. Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York, NY: Oxford University Press.

Freeman, R., and Lewis, R. (1998). *Planning and Implementing Assessment*. London: Kogan Page.

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

Gentry, J. (2015). Has the ODNI Improved U.S. Intelligence Analysis?. *Int. J. Intell. CounterIntell.* 28, 637–661. doi: 10.1080/08850607.2015.1050937

Heazle, M. (2010). Policy lessons from Iraq on managing uncertainty in intelligence assessment: why the strategic/tactical distinction matters. *Intell. Natl Secur.* 25, 290–308. doi: 10.1080/02684527.2010.489780

Intelligence Advanced Research projects Activity [IARPA] (2016). *Broad Agency Announcement (BAA): Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE)*. Washington, DC: Office of Anticipating Surprise.

Intelligence Community Directive [ICD] (2007/2015). *Intelligence Community Directive (ICD) 203, Analytic Standards*. Available at: https://fas.org/irp/dni/icd/icd-203.pdf

IRTPA (2004). Intelligence Reform and Terrorism Prevention Act of 2004. Public Law No 108–458, 118 Stat. 3638.

Lowenthal, M. M. (2012). A disputation on intelligence reform and analysis: my 18 theses. *Int. J. Intell. Counter Intell.* 26, 31–37. doi: 10.1080/08850607.2013.732435

Marchio, J. (2014). Analytic tradecraft and the intelligence community: enduring value, intermittent emphasis. *Intell. Natl. Secur.* 29, 159–183. doi: 10.1080/02684527.2012.746415

Marcoci, A., Vercammen, A., and Burgman, M. (2018). ODNI as an analytic ombudsman: is Intelligence Community Directive 203 up to the task? *Intell. Natl. Secur.* doi: 10.1080/02684527.2018.1546265

Marrin, S. (2012). Evaluating the quality of intelligence analysis: by what (Mis) measure? *Intell. Natl Secur.* 27, 896–912. doi: 10.1080/02684527.2012.699290

National Commission on Terrorist Attacks Upon the United States (2004). *Final Report of the National Commission on Terrorist Attacks Upon the United States*. Washington, DC: USGPO.

O'donovan, B., Price, M., and Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teach. High. Educ.* 9, 325–335. doi: 10.1080/1356251042000216642

Phythian, M. (2006). The perfect intelligence failure? U.S. Pre-war intelligence on iraqi weapons of mass destruction. *Polit. Policy* 34, 400–424. doi: 10.1111/j.1747-1346.2006.00019.x

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Rev. Educ.* 13, 191–209. doi: 10.1080/0305498870130207

Stimson, C., and Habeck, M. (2016). *Reforming Intelligence: A Proposal For Reorganizing the Intelligence Community and Improving Analysis*. Washington, DC: The Heritage Foundation.

Tetlock, P. E., and Mellers, B. A. (2011). Intelligent management of intelligence agencies: beyond accountability ping-pong. *Am. Psychol.* 66, 542–554. doi: 10.1037/a0024285

The Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction (2005). Report to the President of the United States. Washington, DC: USGPO.

van Gelder, T., and de Rozario, R. (2017). "Pursuing Fundamental Advances in Reasoning," in *Proceedings of the Artificial General Intelligence 10 International Conference, AGI-2017. Lecture Notes in Artificial Intelligence*, eds T. Everitt, A. Potapov, and B. Goertzel (Berlin: Springer), 259–262. doi: 10.1007/978-3-319-63703-7_24

Webster, F., Pepper, D., and Jenkins, A. (2000). Assessing the undergraduate dissertation. *Assess. Eval. High. Educ.* 25, 72–80. doi: 10.1080/02602930050025042

Zegart, A. B. (2005). September 11 and the Adaptation Failure of U.S. Intelligence Agencies. *Int. Secur.* 29, 78–111. doi: 10.1162/isec.2005.29.4.78

Zegart, A. B. (2006). An empirical analysis of failed intelligence reforms before September 11. *Polit. Sci. Q.* 121, 33–60. doi: 10.1002/j.1538-165X.2006.tb00564.x

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

PERSPECTIVE
published: 21 December 2018
doi: 10.3389/fpsyg.2018.02640

Correcting Judgment Correctives in National Security Intelligence

David R. Mandel[1]* *and Philip E. Tetlock*[2]

[1] *Intelligence Group, Intelligence, Influence and Collaboration Section, Defence Research and Development Canada, Toronto, ON, Canada,* [2] *Wharton School, University of Pennsylvania, Philadelphia, PA, United States*

Intelligence analysts, like other professionals, form norms that define standards of tradecraft excellence. These norms, however, have evolved in an idiosyncratic manner that reflects the influence of prominent insiders who had keen psychological insights but little appreciation for how to translate those insights into testable hypotheses. The net result is that the prevailing tradecraft norms of best practice are only loosely grounded in the science of judgment and decision-making. The "common sense" of prestigious opinion leaders inside the intelligence community has pre-empted systematic validity testing of the training techniques and judgment aids endorsed by those opinion leaders. Drawing on the scientific literature, we advance hypotheses about how current best practices could well be reducing rather than increasing the quality of analytic products. One set of hypotheses pertain to the failure of tradecraft training to recognize the most basic threat to accuracy: measurement error in the interpretation of the same data and in the communication of interpretations. Another set of hypotheses focuses on the insensitivity of tradecraft training to the risk that issuing broad-brush, one-directional warnings against bias (e.g., over-confidence) will be less likely to encourage self-critical, deliberative cognition than simple response-threshold shifting that yields the mirror-image bias (e.g., under-confidence). Given the magnitude of the consequences of better and worse intelligence analysis flowing to policy-makers, we see a compelling case for greater funding of efforts to test what actually works.

Keywords: judgment and decision making, intelligence analysis, debiasing, error management, corrective action, organizational policies

OPEN ACCESS

Edited by:
Jan B. F. Van Erp,
University of Twente, Netherlands

Reviewed by:
Paul van Schaik,
Teesside University, United Kingdom
Elisabet Tubau,
University of Barcelona, Spain

***Correspondence:**
David R. Mandel
drmandel66@gmail.com

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

Received: *03 October 2018*
Accepted: *10 December 2018*
Published: *21 December 2018*

Citation:
Mandel DR and Tetlock PE (2018) Correcting Judgment Correctives in National Security Intelligence. Front. Psychol. 9:2640. doi: 10.3389/fpsyg.2018.02640

INTRODUCTION

Intelligence organizations in government play a vital role in informing the upper echelons of policymaking, the leaders of nations and their staff who are vested with the responsibility of protecting national security and promoting national interests. Within a given nation, the collective of intelligence organizations – euphemistically known as the intelligence community or, simply, the IC – therefore has an epistemic mandate to deliver timely, relevant, and accurate information to decision makers who operate under time and accountability pressures, the fog of uncertainty, and with foreknowledge that their decisions may alter the course of history.

How then has the IC sought to guarantee for policymakers and the public that they are doing their best to meet their epistemic mandate, given that the vast majority of substantive intelligence relies on human judgments made under conditions of deep uncertainty (Kent, 1964)? Do the IC's tactics to ensure judgment quality rest on sound strategy properly informed by key concepts, methods and findings from judgment and decision science, the field that speaks directly to the challenges the IC faces? To the latter question, we believe the answer is – No. Yet we also

remain optimistic that the IC could substantially improve the quality of its judgments if it took appropriate steps to correct its current corrective strategy – steps that we lay out as a set of IC policy prescriptions.

THE IC'S CURRENT CORRECTIVE APPROACH

The IC is well aware both that its primary analytic product is judgment to support decision-making and that human judgment is prone to bias and error. Sherman Kent, an historian recruited to the fledgling IC during World War II and now widely regarded as the founder of modern intelligence analysis, was keenly concerned about the threats that confirmation bias and groupthink posed to epistemic integrity (Scoblic, 2018). Richards Heuer Jr. went further, documenting in *Psychology of Intelligence Analysis* (Heuer, 1999) how cognitive biases, much of which were revealed in the heuristics-and-biases research program inspired by Kahneman et al. (1982), could skew intelligence judgments and raise the risk of intelligence failure.

Heuer and others improvised simple, back-of-the-napkin, judgment-support methods that analysts could self-apply to debias their judgments and consequently improve their accuracy. The methods, which came to be known as structured analytic techniques or SATs, have proliferated (see Heuer and Pherson, 2014) and continue to represent the IC's main tactical approach to combatting judgment error. In the United States, the Intelligence Reform and Terrorism Prevention Act of 2004 mandated use of SATs and many of them are presented to analysts in intelligence training as methods for coping with their unavoidable "mindsets and biases" (Marchio, 2014; Coulthart, 2017; Chang et al., 2018). More recently, Intelligence Community Directive 203 on analytic standards, promulgated by the Office of the Director of National Intelligence (ODNI), states that analysts "must employ reasoning techniques and practical mechanisms that reveal and mitigate bias" (Office of Director of National Intelligence [ODNI], 2015, p. 2), by which they mean SATs. Variants of this approach have spread to many other nations (e.g., Butler, 2004), an excellent example of a phenomenon that sociologists dub "institutional isomorphism." The SAT paradigm has spread not because there is evidence it works, but because influential professionals in the most powerful organization have endorsed it and no one wants to fall behind prevailing norms of best practices. In these environments, pressures for interoperability can easily trump systematic searches for optimal design, resulting in suboptimal cross-organizational learning.

CRITIQUE OF THE CURRENT APPROACH

The IC's current approach to judgment correctives is flawed for several reasons. We focus here on those that apply to the IC's general approach to judgment correction and do not descend into the weeds to critique individual SATs. Given space constraints, we condense our arguments into two areas of critique: core organizational limitations and core conceptual limitations. These areas are related, and have a common denominator in the IC's slow uptake from judgment and decision science, which followed from its commitment to an incidental approach, or lack of interest in pursuing a sustained, programmatic, and scientific approach to tradecraft innovation. We briefly address that common denominator before turning to the two areas of critique.

The Incidental Approach to IC Innovation

The IC's current approach to judgment correctives emerged from the attention of a handful of diligent analysts to specific problems they encountered in the practice of intelligence from the 1940s to 1980s. For instance, Kent's stubborn preoccupation with improving the fidelity of communications of uncertainty estimates was affected by his direct experience with a policy-maker who was unsure of the meaning of the expression, "serious possibility," that appeared in a 1951 National Intelligence Estimate on the probability of a Soviet invasion of Yugoslavia that year (Kent, 1964). When Kent asked his colleagues on the Board of National Estimates what they thought the term meant, he got answers ranging from 1:4 to 4:1 odds, which Kent described as jolting. Similarly, Heuer's interest in intelligence tradecraft – and "alternative analysis," in particular – was sparked by his involvement in the case of Soviet KGB defector Yuri Nosenko and his conclusion that the United States IC made inadequate effort to consider alternative explanations for a string of suspicious events that seemed to support the conclusion that Nosenko was a KGB disinformation agent (Heuer, 1987).

These tradecraft mavericks deserve credit for their trailblazing efforts to improve the practice of intelligence analysis. However, their examples also lay bare the adverse consequences of an *ad hoc*, character-driven approach to developing tradecraft. Critically, none of these tradecraft developers had advanced expertise in judgment and decision science. For example, although Heuer was well read in literature on higher-order cognition, he did not pursue it at a professional or even post-graduate level, and he was not trained in research methods and statistical analysis. It is therefore unsurprising that he did not subject his methods – notably the Analysis of Competing Hypotheses (ACH) technique – to experimental tests of whether they actually improved judgment in measurable ways.

Organizational Limitations

Testing hypotheses is fundamental to both basic and applied sciences. Even our best ideas need to be put to rigorous empirical tests because most good ideas still fail. Mandel (in press) recently argued that the IC's approach to tradecraft development follows what he called the *goodness heuristic*. Using this heuristic, if, upon mental inspection, an idea such as an imagined SAT for debiasing judgment seems good, then one should act on it as if it were in fact good because it probably is good. The goodness heuristic, which rests on a very likely excessively optimistic prior probability for ideational success, therefore takes Kahneman's (2011) WYSIATI (what-you-see-is-all-there-is) principle to the next level by elucidating its implications for action by individuals and organizations.

Yet, as any seasoned scientist knows, not only do good ideas need to be rigorously tested, they need to be tested using multi-task and multi-benchmark methods (e.g., Mellers et al., 2017). There also should ideally be a diverse pool of ideas being tested by independent clusters of researchers, and among those clusters there must be a healthy sense of competition in epistemic tournaments, whether organized or *ad hoc* (e.g., Tetlock et al., 2017). This is vital because scientists, as theorists, can become prisoners of their preconceptions all too easily (Tetlock and Henik, 2005). Moreover, scientists, like all individuals, pursue goals other than purely epistemic ones (Mandel and Tetlock, 2016). It is vital, therefore, that scientists' ideas and key findings be subject to peer scrutiny.

Those who shaped the IC's current approach to judgment correctives varied in their commitment to testing ideas scientifically. Heuer, who had the greatest direct impact on the SAT approach to judgment correctives, questioned the value of science in adjudicating on the merits of proposed corrective methods. In an August 15, 2010 response to suggestions posted on an online discussion of the International Association for Intelligence Education that his ACH technique be empirically tested, Heuer wrote:

> Can't we have confidence in making a common sense judgment that going through the process of assessing the inconsistency of evidence will generally improve the quality of analysis? Similarly, can't we have confidence in making a common sense judgment that starting the analysis with a set of hypotheses will, on average, lead to better analysis than starting by looking at the pros and cons for a single hypothesis? Do we really need an empirical analysis of these two points? Is it really feasible to do a high quality empirical analysis of the effectiveness of these two points?[1]

He also expressed reservations about the feasibility of experiments to test methods such as ACH, concluding, "If the empirical testing of my two claims about the value of ACH doesn't replicate exactly how ACH is (or should be) used in the Intel Community, I would be inclined to ignore it and stick with my common sense judgment."

It is ironic that one of the IC's foremost tradecraft contributors, who stressed the importance of combatting confirmation bias, would take this stand. Yet the inconsistency should not shock us. The double standard – intuition is fine for me, but not for you – is simply more anecdotal evidence of the well-documented *bias blind spot*, the tendency to perceive biases in others' thinking and judgments more easily than in one's own (Pronin et al., 2002).

We do not blame Heuer and others for exhibiting what most of us exhibit to varying degrees, but his stance highlights a consequence of the IC's decision over much of its history to invest very little in improving judgment quality through science, while investing heavily in collections technology. Over the last decade, the United States IC has changed this approach and now funds the Intelligence Advanced Research Projects Activity (IARPA), which is programmatic, engaging large numbers of scientists from industry and academia, and which has led to important scientific advances that hold promise for improving intelligence products. Whether these advances can be effectively integrated into the analytic training and workflows of intelligence organizations remains to be seen.

[1]Heuer, R. J., Jr., August 15, 2010 email correspondence sent to the International Association for Intelligence Education.

Conceptual Limitations

The IC's traditional approach to analytic tradecraft has also fostered conceptual setbacks. While a heavy emphasis is placed on the mitigation of cognitive biases, virtually no attention is given to the problem of imprecision and unreliability caused by "noisy" unsystematic error (Chang et al., 2018). Moreover, cognitive biases are conceptualized as unipolar phenomena needing to be reduced rather than as bipolar phenomena in which bias reduction strategies would require knowing where one was starting from, both in terms of direction and magnitude. Consequently, undue faith has been placed in assumptions regarding what types of biases needed to be corrected. For instance, whereas overconfidence is seen as problematic and attention is drawn to it in analytic training, the polar-opposite bias, underconfidence, is virtually ignored. However, recent studies show evidence of underconfidence in strategic intelligence forecasts (Mandel and Barnes, 2014, 2018) and in intelligence analysts' probability judgments in experimental tasks (Mandel, 2015).

When we look at the research literature on how people cope with accountability demands (Lerner and Tetlock, 1999), we worry that the IC's indiscriminate injunctions to beware of overconfidence will mainly yield indiscriminate response-threshold shifts – and the mirror-image bias of underconfidence. The net effect will be to further water down the informativeness of intelligence assessments for decision makers with excessive uncertainty. Similarly, the main effect of broad-brush warnings about confirmation bias might well be to induce endless second-guessing, to the point of analysis paralysis. Ultimately, the unipolar view of cognitive bias has allowed the IC to conveniently skirt value-laden, vexing questions about how bias-reduction tradeoffs should be resolved.

The IC's error-neglect blind spot is equally troubling. Not only has the IC not taken proactive measures to minimize noise in intelligence judgments, noise neglect signals that the IC has not carefully considered how the very techniques they promote to minimize bias might amplify noise (Chang et al., 2018). Yet the weakly defined multistep processes that most SATs represent are no less than covert greenhouses for noise production. While giving the appearance of a standardized judgment-support process, SATs actually leave a long list of implementation decisions to analysts. How much agreement is there among analysts on such decisions? How reliably do the same analysts make these decisions over time? The few extant studies do not inspire optimism. For example, analysts asked to judge the probability of information accuracy on the basis of Admiralty-code ratings of source reliability (i.e., A–F) and information credibility (i.e., 1–6) were unreliable when the two ratings were incongruent in ordinal value, and inter-analyst agreement plummeted as scale incongruence increased (Mandel, 2018, Annex D).

In comparison to the Admiralty code, SATs like ACH create vast opportunities for inconsistency to flourish. To take just one example, consider the engine of ACH, which involves listing evidence in rows, hypotheses in columns, and then assessing the degree of consistency in each cell of the matrix. The meaning of consistency is left up to the analyst to interpret. One might treat it as the probability of the evidence given the hypothesis, while another might treat it as the inverse of that probability. Another still might assess whether the hypothesis necessarily follows from the evidence or vice versa, while yet another might run the test but with plausibility substituting for necessity. Perhaps the most common approach is to judge the representativeness of one to the other. In that case, and not without a touch of irony, ACH would be promoting the use of the representativeness heuristic under the guise of a debiasing strategy.

CORRECTING THE IC'S CURRENT CORRECTIVE APPROACH

Both the organizational and conceptual limitations of the IC's approach to judgment correctives, in particular, and analytic tradecraft, in general, stem from its *ad-hoc*, unscientific and character-driven nature. For the IC to develop effective correctives, it should abandon the complacent strategy of waiting for the next Kent or Heuer to spontaneously arise. The IC needs a diverse infusion of ideas from scientists outside the IC. It needs those scientists not only to put forward their best ideas, but also to test them in rigorous experiments or experimental tournaments. The IC should take the most promising results and work with scientific teams to transition the ideas into analytic processes. Those teams should also work with their IC counterparts to devise rigorous ways of trialing those processes, and the results of those trials should be taken seriously. What might work in an IARPA tournament, might not work so well in practice. If not, then reasons for variance in efficacy should be examined. Is the original idea doomed to transition failure, or was the transition strategy flawed but correctable?

The IC also should abandon the assumption that analytic judgments made in the absence of SATs must be intuitive and flawed. They should further banish the corollary view that although a SAT might not be perfect, it's better than nothing. The first assumption is certainly wrong and the second is probably wrong too. While intuitive processes enter into analysts' judgments, surely so can deliberative thought. SATs foster the illusion that intuition is driven from the judgment process. In reality, it is likely transferred to the process of conducting the SAT exercise itself. The effects of such transfer can be far from banal. For instance, SATs might disrupt good deliberative reasoning about the substantive issues. They might bolster undeserved confidence in the accuracy and logical coherence of analysts' judgments. And they might foster IC complacency through the belief that corrective measures are sound and sufficient. For example, Mandel et al. (2018) report that intelligence analysts who were trained in ACH and who were instructed to use ACH to solve a probabilistic hypothesis-testing task were significantly more susceptible to coherence-violating unpacking effects (Tversky and Koehler, 1994) than a control sample of analysts from the same cohort who were not trained in ACH and who were left to their own reasoning devices.

Finally, the IC should broaden its horizons and start thinking beyond the analyst. All SATs share a focus on supporting the analyst, whether individually or in teams. Yet no attention has been given to how intelligence organizations might improve the accuracy of assessments through a range of post-analytic means such as recalibrating probabilistic judgments to correct for observable biases and aggregating judgments to boost signal-to-noise ratios through error cancelation and performance-sniffing methods. Recalibrating forecasts to make them more extreme has been shown to improve calibration in IARPA's "ACE" geopolitical forecasting tournament (Baron et al., 2014; Turner et al., 2014) and in actual strategic intelligence forecasts (Mandel and Barnes, 2014). Likewise, recalibration methods that "coherentize" probability judgments by forcing them to respect one or more axioms of probability calculus, such as additivity and unitarity, can improve accuracy (Karvetski et al., 2013). The IC could also leverage decades of research on the benefits of statistically aggregating probability estimates. Taking an unweighted arithmetic average of multiple estimates is a highly effective method of error cancelation (Clemen and Winkler, 1999). More sophisticated aggregation methods that exploit individual differences in coherence (Predd et al., 2008; Wang et al., 2011; Karvetski et al., 2013) or other measurable aspects of performance (Cooke and Goossens, 2008) also hold promise for the IC. Indeed, Mandel et al. (2018) found that analysts' judgment accuracy was substantially improved by first coherentizing and then aggregating their judgments.

To accelerate the discovery process, the IC should also take steps to systematically monitor the accuracy of its products. This will reveal the types of corrective actions most needed, and it can also shed light on factors that predict judgment accuracy. The results may be counter-intuitive and impossible to predict from theory. For instance, contrary to intuitive expectation, topic-related expertise among cancer research experts did not predict better accuracy in forecasting the reproducibility of cancer trial results, but expertise defined in terms of publication impact (h-index) did (Benjamin et al., 2017). Likewise, Tetlock (2005) found that political experts working inside their self-described domain of competence were no more accurate than experts working outside their domain in a geopolitical forecasting tournament. Ferreting out the factors that could be used in performance-sniffing weighting methods will take time and research effort, but these and other post-analytic interventions could significantly boost the IC's judgment accuracy in years to come. The IC only needs to reduce the probability of a trillion-dollar mistake by a tiny amount to justify multi-million-dollar research investments.

AUTHOR CONTRIBUTIONS

Both authors contributed to the thinking behind and writing of this article.

FUNDING

Funding support for this work provided by the Canadian Safety and Security Program projects CSSP-2016-TI-2224 (Improving Intelligence Assessment Processes with Decision Science) and CSSP-2018-TI-2394 (Decision Science for Superior Intelligence Production), and Department of National Defence project 05da (Joint Intelligence Collection and Analytic Capability).

REFERENCES

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., and Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decis. Anal.* 11, 133–145. doi: 10.1287/deca.2014.0293

Benjamin, D., Mandel, D. R., and Kimmelman, J. (2017). Can cancer researchers accurately judge whether preclinical reports will reproduce? *PLoS Biol.* 15:e2002212. doi: 10.1371/journal.pbio.2002212

Butler, L. (2004). *Review of Intelligence on Weapons of Mass Destruction: Report of a Committee of Privy Councillors*. London: The Stationery Office.

Chang, W., Berdini, E., Mandel, D. R., and Tetlock, P. E. (2018). Restructuring structured analytic techniques in intelligence. *Intell. Natl. Secur.* 33, 337–356. doi: 10.1080/02684527.2017.1400230

Clemen, R. T., and Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Anal.* 19, 187–203. doi: 10.1111/j.1539-6924.1999.tb00399.x

Cooke, R. M., and Goossens, L. L. H. J. (2008). TU Delft expert judgment data base. *Reliabil. Eng. Syst. Saf.* 93, 657–674. doi: 10.1016/j.ress.2007.03.005

Coulthart, S. J. (2017). An evidence-based evaluation of 12 core structured analytic techniques. *Int. J. Intell. CounterIntell.* 30, 368–391. doi: 10.1080/08850607.2016.1230706

Heuer, R. J. Jr. (1987). Nosenko: five paths to judgment. *Stud. Intell.* 31, 71–101.

Heuer, R. J. Jr. (1999). *Psychology of Intelligence Analysis*. Washington, DC: Center for the Study of Intelligence.

Heuer, R. J. Jr., and Pherson, R. H. (2014). *Structured Analytic Techniques for Intelligence Analysis*. Washington, DC: CQ Press.

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

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511809477

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," in *Sherman Kent and the Board of National Estimates: Collected Essays*, ed. D. P. Steury (Washington, DC: Center for the Study of Intelligence), 133–146.

Lerner, J. S., and Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychol. Bull.* 125, 255–275. doi: 10.1037/0033-2909.125.2.255

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

Mandel, D. R. (2018). "Annex g: report on sas-114 experiment on analysis of competing hypotheses," *in Proceedings of the SAS-114 Workshop on Communicating Uncertainty, Assessing Information Quality and Risk, and Using Structured Techniques in Intelligence Analysis*, ed. D. R. Mandel (Brussels: NATO STO), doi: 10.14339/STO-MP-SAS-114

Mandel, D. R. (in press). "Can decision science improve intelligence analysis?," in *Correcting Judgment Correctives in Intelligence: A Reader*, eds S. Coulthart, M. Landon-Murray, and D. Van Puyvelde (Washington, DC: Georgetown University Press).

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

Mandel, D. R., and Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *J. Behav. Decis. Mak.* 31, 127–137. doi: 10.1002/bdm.2055

Mandel, D. R., Karvetski, C. W., and Dhami, M. K. (2018). Boosting intelligence analysts' judgment accuracy: what works, what fails? *Judgm. Decis. Mak.* 13, 607–621.

Mandel, D. R., and Tetlock, P. E. (2016). Debunking the myth of value-neutral virginity: toward truth in scientific advertising. *Front. Psychol.* 7:451. doi: 10.3389/fpsyg.2016.00451

Marchio, J. (2014). Analytic tradecraft and the intelligence community: enduring value, intermittent emphasis. *Intell. Natl. Secur.* 29, 159–183. doi: 10.1080/02684527.2012.746415

Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., and Tetlock, P. E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgm. Decis. Mak.* 12, 369–381.

Office of Director of National Intelligence [ODNI] (2015). *Intelligence Community Directive 203: Analytic Standards*. Washington, DC: Office of Director of National Intelligence.

Predd, J. B., Osherson, D. N., Kulkarni, S. R., and Poor, H. V. (2008). Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decis. Anal.* 5, 177–189. doi: 10.1287/deca.1080.0119

Pronin, E., Lin, D. Y., and Ross, L. (2002). The bias blind spot: perceptions of bias in self versus others. *Pers. Soc. Psychol. Bull.* 28, 369–381. doi: 10.1177/0146167202286008

Scoblic, J. P. (2018). *Beacon and Warning: Sherman Kent, Scientific Hubris, and the Cia's Office of National Estimates. Texas National Security Review*. Available at: https://tnsr.org/2018/08/beacon-and-warning-sherman-kent-scientific-hubris-and-the-cias-office-of-national-estimates/.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?*. Princeton, NJ: Princeton University Press.

Tetlock, P. E., and Henik, E. (2005). "Theory- versus imagination-driven thinking about historical counterfactuals: are we prisoners of our preconceptions?," in *The Psychology of Counterfactual Thinking*, eds D. R. Mandel, D. J. Hilton, and P. Catellani (New York, NY: Routledge).

Tetlock, P. E., Mellers, B. A., and Scoblic, J. P. (2017). Bringing probability judgments into policy debates via forecasting tournaments. *Science* 355, 481–483. doi: 10.1126/science.aal3147

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., and Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Mach. Learn.* 95, 261–289. doi: 10.1007/s10994-013-5401-4

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

Wang, G., Kulkarni, S. R., Poor, H. V., and Osherson, D. N. (2011). Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decis. Anal.* 8, 128–144. doi: 10.1287/deca.1110.0206

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership