Cognition 245 (2024) 105722

Contents lists available at ScienceDirect

Cognition

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

When intuitive Bayesians need to be good readers: The problem-wording effect on Bayesian reasoning

Miroslav Sirota [a,*], Gorka Navarrete [b], Marie Juanchich [a]

[a] Department of Psychology, University of Essex, United Kingdom
[b] Center for Social and Cognitive Neuroscience (CSCN), School of Psychology, Universidad Adolfo Ibáñez, Santiago de Chile, Chile

ARTICLE INFO

Keywords:
Bayesian reasoning
Natural frequencies
Wording effect
Mathematical problem-solving

ABSTRACT

Are humans intuitive Bayesians? It depends. People seem to be Bayesians when updating probabilities from experience but not when acquiring probabilities from descriptions (i.e., Bayesian textbook problems). Decades of research on textbook problems have focused on how the format of the statistical information (e.g., the natural frequency effect) affects such reasoning. However, it pays much less attention to the wording of these problems. Mathematical problem-solving literature indicates that wording is critical for performance. Wording effects (the wording varied across the problems and manipulations) can also have far-reaching consequences. These may have confounded between-format comparisons and moderated within-format variability in prior research. Therefore, across seven experiments ($N = 4909$), we investigated the impact of the wording of medical screening problems and statistical formats on Bayesian reasoning in a general adult population. Participants generated more Bayesian answers with natural frequencies than with single-event probabilities, but only with the improved wording. The improved wording of the natural frequencies consistently led to more Bayesian answers than the natural frequencies with standard wording. The improved wording effect occurred mainly due to a more efficient description of the statistical information—cueing required mathematical operations, an unambiguous association of numbers with their reference class and verbal simplification. The wording effect extends the current theoretical explanations of Bayesian reasoning and bears methodological and practical implications. Ultimately, even intuitive Bayesians must be good readers when solving Bayesian textbook problems.

Are naïve people intuitive Bayesians? If so, it was argued that Bayesian inference can serve as a basis for psychological models accounting for human cognition and behaviour (Peterson & Beach, 1967). This question attracted decades of dedicated research (e.g., Bar-Hillel, 1980; Cosmides & Tooby, 1996; Edwards, Lindman, & Phillips, 1965; Kahneman & Tversky, 1972; Peterson & Beach, 1967). This research arrived at profoundly contradictory conclusions. On the one hand, people were intuitive Bayesians when updating probabilities from their experience (Armstrong & Spaniol, 2017; Cohen, Sidlowski, & Staub, 2017; Edwards et al., 1965; Griffiths & Tenenbaum, 2006; Peterson & Beach, 1967; Vallee-Tourangeau, Abadie, & Vallee-Tourangeau, 2015). For example, people who estimated the posterior probability of having a rare disease, such as insulin-dependent diabetes, learnt from a sequential presentation of representative cases. On the other hand, people seemed to be biased and departed substantially from Bayesian normative responses when updating probabilities from a description, i.e. verbal statistical summaries, also known as Bayesian textbook problems (Armstrong & Spaniol, 2017; Bar-Hillel, 1980; Kahneman & Tversky, 1972, 1973). For example, people found it challenging to correctly calculate the posterior probability of having a rare disease, such as insulin-dependent diabetes, if a diagnostic test was positive (Gigerenzer & Hoffrage, 1995; Hoffrage & Gigerenzer, 2004; Siegrist & Keller, 2011).

How can we explain this contradiction? It was recently suggested that the experimental protocol differences could account for the contradictory findings (Lejarraga & Hertwig, 2021; Sirota, Vallee-Tourangeau, Vallee-Tourangeau, & Juanchich, 2015). In the experimental approach of the "people as intuitive Bayesians" research programme, participants experienced the statistical information, saw stimuli presented sequentially and were offered feedback—they learnt from experience. In contrast, in the experimental approach of the "heuristics and biases" research programme, people were typically asked to solve word problems (e.g., Bayesian textbook problems). Participants read about the *statistical information* presented in an aggregated

* Corresponding author.
E-mail address: msirota@essex.ac.uk (M. Sirota).

https://doi.org/10.1016/j.cognition.2024.105722
Received 15 June 2023; Received in revised form 30 November 2023; Accepted 12 January 2024
Available online 2 February 2024

form embedded in a *written description*. In other words, people who make intuitive Bayesian inferences from experience may find it hard to solve Bayesian word problems—a finding that was recently experimentally demonstrated (Armstrong & Spaniol, 2017; Schulze & Hertwig, 2021).

Much research was devoted to understanding how the format of the aggregated statistical information can facilitate Bayesian problem-solving (Gigerenzer & Hoffrage, 1995; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000; for a meta-analysis, see McDowell & Jacobs, 2017). However, surprisingly little research focused on understanding the role of *the written description of the problem—how the problems are worded*. Yet the use of verbal descriptions seemed to be one of the critical differential elements in the two research programmes. Since little research attention was paid to this wording, it often varied between and within statistical format manipulation, therefore potentially confounding between-format comparisons and moderating within-format variability in prior research. The task wording might therefore bear critical theoretical and methodological implications. In addition, if, as expected, the wording shapes Bayesian reasoning performance, it will have important practical significance. This is because adequately estimating conditional probabilities from descriptions remains vital for informed decision-making in essential domains of life, such as health care (Navarrete, Correia, & Froimovitch, 2014). Therefore, in this manuscript, we aimed to better understand the effect of wording on Bayesian reasoning in textbook problems.

1. Format effects-centered research and theory-building

Prior research on reasoning with Bayesian textbook problems predominantly focused on the format of aggregated statistical information. The pivot finding in this literature remains the facilitative effect of natural frequencies relative to conditional probabilities on Bayesian reasoning (Gigerenzer & Hoffrage, 1995; McDowell & Jacobs, 2017; Siegrist & Keller, 2011). Natural frequencies represent frequencies with a natural sampling structure encapsulating the sequential encoding of events. (E.g., "10 women with positive tests out of 1,000 women taking the test" and "2 women with cancer out of the 10 women with a positive test".) Natural frequencies are often compared with conditional probabilities encapsulating single-event probability with a normalised structure. (E.g., "The probability of having a positive test is 1%." And, "The probability of having cancer if you receive a positive test is 20%.") Indeed, a recent meta-analysis aggregating findings across populations, scenarios and contexts showed robust evidence that natural frequencies facilitated Bayesian performance compared with conditional probabilities (McDowell & Jacobs, 2017).

The research concerning the natural frequency effect became so central that almost all theory-building and testing related to Bayesian reasoning concentrated on the explanation of this effect. (e.g., Barbey & Sloman, 2007; Brase & Hill, 2017; Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999; Lesage, Navarrete, & De Neys, 2013; Pighin, Girotto, & Tentori, 2017; Sedlmeier & Gigerenzer, 2001; Sirota, Juanchich, & Hagmayer, 2014; Sloman, Over, Slovak, & Stibel, 2003) Two dominant sets of theories emerged. On the one hand, ecological rationality theories explained the facilitating effect of natural frequencies as a result of the ecological fit between natural frequencies and the way humans encountered information for thousands of years during human evolution (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Hoffrage et al., 2000). Some authors even proposed stronger claims that we evolved a domain-specific cognitive mechanism operating on frequencies (Brase & Hill, 2017; Cosmides & Tooby, 1996). On the other hand, nested set theories explained the relative success of natural frequencies as reasoning over a problem with statistical information organised in a salient, nested set structure (e.g., Barbey & Sloman, 2007; Girotto & Gonzalez, 2001; Pighin, Tentori, & Girotto, 2017; Sloman et al., 2003; Tversky & Kahneman, 1983). Thus, the formats, such as chances with natural sampling, could equally improve Bayesian reasoning because they used the nested set structure, even though they did not feature frequencies (Girotto & Gonzalez, 2001). The central role of the format effects underlined the fact that it leaked into the theoretical explanations of other effects. These included visual aids effects (e.g., Yamagishi, 2003), training effects (e.g., Sedlmeier & Gigerenzer, 2001; Sirota, Kostovicova, & Vallee-Tourangeau, 2015) and the role of individual differences (e.g., Brase & Hill, 2017; Chapman & Liu, 2009; Sirota, Juanchich, & Hagmayer, 2014). For instance, the mechanisms behind the beneficial effect of visual representations of information (e.g., icon arrays) were often predicted and interpreted using the two theoretical frameworks. Icons were considered to be individualised representations tapping more effectively into the frequency encoding mechanism (Brase & Hill, 2017; Sirota, Kostovicova, & Juanchich, 2014; Yamagishi, 2003).

Notwithstanding the importance of the format effects in Bayesian reasoning, the meta-analysis showed a substantial variation in the strength of the natural frequency effect and a wide variation in the levels of absolute normative performance within the natural frequency format (McDowell & Jacobs, 2017). For instance, pregnant women, their companions and midwives did not benefit from the effect of natural frequencies when solving highly relevant screening problems on the risk of an unborn child having Down syndrome. Their performance was generally low, with midwives giving zero correct answers (Bramwell, West, & Salmon, 2006). Indeed, the facilitative effect of natural frequencies could be very limited with medical scenarios tested in a general adult population (Pighin, Gonzalez, Savadori, & Girotto, 2016; Siegrist & Keller, 2011)—but see counter-evidence (McDowell, Galesic, & Gigerenzer, 2018). Conversely, almost 90% of highly numerate university students correctly solved a short, simply-worded, green and red apples Bayesian problem (Johnson & Tubau, 2013). In addition, even within the natural frequency format, the same meta-analysis revealed a substantial variation in the proportion of correct responses across tasks and studies. Several moderators were identified, such as problem representation (e.g., short menu, question form) and methodological moderator (e.g., scoring criteria). However, these did not account for all the variability observed, which could have been because of problem-wording variations (McDowell & Jacobs, 2017).

To summarise, prior research on Bayesian reasoning based on textbook problems mostly focused on the effect of statistical formats, but the wording of the problems was not systematically investigated. Still, the wording of the problems differed between the statistical formats (e.g., natural frequency vs conditional probability). It varied within the format of the natural frequencies themselves as the wording of the problems was not standardised. And, finally, it remains a critical differential element of the Bayesian reasoning from experience and descriptions.

1.1. Advancing the debate: problem-wording affects Bayesian problem-solving

It is not just a lack of focus on the problem-wording, despite its potential theoretical and methodological consequences; there are good reasons to believe that the problem-wording will likely affect Bayesian problem-solving. First, some limited evidence showed that specific changes to the problem-wording affected Bayesian reasoning. For instance, participants were asked to solve the problem using the same sample in the question as presented in the problem rather than a "similar new sample"—so the samples introduced in the problem and the question were identical. In that instance, the same type of sample increased substantially the proportion of correct Bayesian responses (Johnson & Tubau, 2017). Furthermore, clarifying the causal model of the problem was also shown to yield, even though somewhat mixed, evidence of better normative performance (Krynski & Tenenbaum, 2007; McNair & Feeney, 2014, 2015). For example, providing an explicit alternative cause for false positive tests (e.g., a benign cyst explaining a positive mammogram) improved Bayesian reasoning in the mammography problem (Krynski & Tenenbaum, 2007).

Moreover, reducing the verbal complexity of the word problems—making the word problem shorter by stripping down unnecessary information—improved Bayesian reasoning with natural frequencies and conditional probability problems relative to the wordier versions of the same problems (Johnson & Tubau, 2013). Finally, no facilitatory benefit of natural frequencies was observed with equivalent wording by introducing the textual partitive structure to the problems featuring normalised frequencies (Macchi, 2000). However, this latter finding might have an alternative explanation since the base rate information was provided as a frequency ratio (e.g., 10 out of 100), which might have encouraged participants to translate the normalised frequencies into natural frequencies.

Second, a plethora of studies in mathematical problem-solving literature stresses the importance of problem-wording and verbal skills required for text comprehension in successful mathematical problem-solving. (e.g., Fuchs, Fuchs, Compton, Hamlett, & Wang, 2015; Glenberg, Willford, Gibson, Goldberg, & Zhu, 2012; Hadianto et al., 2020; Hegarty, Mayer, & Monk, 1995; Leiss, Plath, & Schwippert, 2019; Strohmaier et al., 2022; Vilenius-Tuohimaa, Aunola, & Nurmi, 2008; Zhou et al., 2018) For instance, reading comprehension was the best predictor of successfully solving complex mathematical word problems in adults (e.g., Strohmaier et al., 2022). Neuroimaging evidence indicated that the semantic network in the brain responsible for language processing was more activated when adults were solving mathematical word problems than equivalent arithmetic computations (Zhou et al., 2018). Similarly, reading comprehension reliably predicted children's ability to solve word problems (e.g., Leiss et al., 2019; Vilenius-Tuohimaa et al., 2008). In addition, experimental studies in children indicated that small changes in how the problem was worded and contextually presented substantially affected accuracy (e.g., Cummins, 1991; Cummins, Kintsch, Reusser, & Weimer, 1988; Davis-Dorsey, Ross, & Morrison, 1991; Staub & Reusser, 1995). These changes included presenting strong action language, avoiding misunderstanding of quantification or reference (e.g., "altogether" interpreted as meaning "each"), using an explicit set reference language (e.g., "the rest of the group"), familiar situations and presenting the problem in chronological order. These amendments facilitated mapping the problem text onto the mathematical relationships, and, in turn, improved problem-solving accuracy. (see an overview in LeBlanc & WeberRussell, 1996).

In addition to this corpus of research evidence, various theoretical frameworks from the mathematical problem-solving literature—such as the construction-integration model, the semantic congruence theory and other computational models—suggested that the way a problem was worded would affect problem-solving performance. (e.g., Gros, Thibaut, & Sander, 2020; Kintsch, 1988; Kintsch & Greeno, 1985; LeBlanc & WeberRussell, 1996) For instance, according to the construction-integration model, mathematical problem-solving can be described as a bottom-up process of construction and integration processes to form an appropriate problem model. In the construction process, a person constructs a network of activated concepts and propositions—text base—by transforming the linguistic input into conceptual representations and their meaning and a real-world knowledge associative network. The resulting text base is enriched but also incoherent and potentially contradictory. The integration process then forms the text base into a coherent text representation by excluding the elements with low activation. Thus, this process model describes in detail how reading comprehension and knowledge of the real world help to build an appropriate problem model, which leads to the successful solution of arithmetic word problems (Kintsch, 1988).

These solid, empirical and theoretical reasons made us believe that the wording, linguistic and presentational issues would affect the successful solution rate of Bayesian textbook problems similarly as these issues affected problem-solving in other areas of mathematical problem-solving.

1.2. Present research

In the present research, we aimed to test the effect of problem-wording on Bayesian reasoning in a general adult population. Guided by the literature on mathematical problem-solving outlined above, we reworded some Bayesian textbook problems to enhance problem representation and, in turn, facilitate Bayesian reasoning (see Table 1). In other words, while keeping the statistical format constant, we changed the way the statistical information was verbally described. We used medical screening scenarios from prior research (Galesic, Gigerenzer, & Straubinger, 2009; Pighin et al., 2016); we considered these wordings standard. In the improved wording, we ensured that the numerical information was unambiguously linked with the reference categories. E.g., "12 of every 15 such women" was replaced with "12 [reference category] out of the 15 [reference category]", to make encoding information easier and less confusing. When introducing false positive results, we added the phrase: "In addition to those 12 women" to further ensure better text integration using an explicit reference to a previously described set, while cueing the required mathematical operations (i.e., addition). Furthermore, we made the sentences simpler to follow by removing conditional sentences. We also reworded the question so that the order of categories matched the response ratio (e.g., the number of women with a child with Down syndrome out of the women who got received a positive result).

The proposed wording changes were designed to test the idea that wording matters—they represented improved wording rather than the optimised wording determined via an iterative design. We ran seven experiments to test the wording effect of natural frequencies in problems featuring natural frequencies. We assessed the wording effect across

Table 1
Standard and improved wording of Bayesian textbook problem (Down Syndrome Problem).

Standard Wording	Improved Wording
To determine whether an unborn child has Down syndrome, doctors sometimes measure the thickness of the fetus' neck skin fold.	To determine whether an unborn child has Down syndrome, doctors sometimes measure the thickness of the fetus' neck skin fold.
Here is some information about that 'neck-fold' test.	Here is some information about that 'neck-fold' test.
• 15 out of every 10,000 pregnant women are pregnant with a child who has Down syndrome. • When a woman is pregnant with a child that has Down syndrome, it is not sure that she will have a positive result on the 'neck-fold' test. Specifically, 12 of every 15 such women will have a positive result on the 'neck-fold' test. • When a woman is pregnant with a child that does not have Down syndrome, it is still possible that she will get a positive result on the 'neck-fold' test. Specifically, 799 out of every 9985 of such women will have a positive result on the 'neck-fold' test.	• 15 out of every 10,000 pregnant women are pregnant with a child who has Down syndrome. • **12 women will have a positive result on the 'neck-fold' test out of the 15 women pregnant with a child that has Down syndrome.** • **In addition to those 12 women, 799 women will also have a positive result on the 'neck-fold' test out of the remaining 9985 women pregnant with a child that does not have Down syndrome.**
Here is a new representative sample of pregnant women who got a positive result on the 'neck-fold' test. Please estimate how many of these women do you expect to have a child with Down syndrome.	Here is a new representative sample of pregnant women who got a positive result on the 'neck-fold' test. **Please estimate how many women actually have a child with Down syndrome out of the women who got a positive result on the 'neck-fold' test.**

The standard wording was taken from prior research (Galesic et al., 2009; Pighin et al., 2016). The improved wording (bolded) consisted of rewording statistical information (i.e., the close association of the numbers with the reference categories, clear cueing mathematical operations and simplifying the text). The question was also reworded to match the order of the presented information with the order presented in the response question (X out of Y).

different screening problems, numerical values and statistical formats (Experiments 1–7). We tested the relative and additive effect of improved wording with other problem representation enhancement strategies: the same sample type and causal explanation (Experiments 2 and 3). We also decomposed the wording effect by testing the critical components of the problem's rewording (Experiment 4). The overarching hypothesis was that the improved wording of the problem would increase the proportion of Bayesian responses compared with the standard wording of the textbook problems used in prior research.

Moreover, we used a general adult population to revisit the existence and strength of three effects reported in the Bayesian reasoning literature typically studied with undergraduate students. First, we tested the format effect of statistical information—natural frequencies compared with single-event probability and normalised frequencies—on Bayesian reasoning (Experiments 1, 3, 6, and 7). Given the prior literature, we expected that the natural frequency formats would increase the proportion of correct Bayesian answers compared with the normalised formats. Second, we tested the effects of two problem representation changes that have been shown to facilitate Bayesian reasoning: the same sample type and causal explanation (Experiments 2 and 3). We expected that both representational changes would increase the proportion of Bayesian responses compared with the standard wording of the textbook problem.

1.3. Open science statement

We conducted all studies presented in this manuscript in accordance with the ethical standards of the American Psychological Association (APA) and obtained ethical approval from the institution of the first author. We have reported how we determined our sample sizes, all measures, manipulations and exclusions in all experiments. Experiments 1–3 were not preregistered; Experiments 4–7 were preregistered (see https://aspredicted.org/AGA_XCA, https://aspredicted.org/T8D_7GH, https://aspredicted.org/GYB_HL8, https://aspredicted.org/JTS_ZM6). The materials, data sets with the codebook, R code for statistical analyses, and preregistrations are publicly available on the Open Science Framework at osf.io/kp3g7.

2. Experiment 1

We followed two main aims in this experiment. First, we aimed to estimate the effect of improved, relative to standard, wording of the problem featuring the statistical natural frequency format. Second, we aimed to test the facilitative effect of natural frequencies, relative to the single-event probability format, using a medical screening test in a general adult population. We used a verbally complex and realistic problem from prior research that was previously used to test the facilitative effect of natural frequencies (Galesic et al., 2009; Pighin et al., 2016). Our experiment would enable us to disentangle the effect of wording from the effect of statistical format. We hypothesised that natural frequencies in standard wording and improved wording would facilitate Bayesian reasoning relative to the single-event probability format. More importantly, we hypothesised that the improved wording would have an additional facilitative effect for problems featuring natural frequencies.

2.1. Method

2.1.1. Participants and design

We aimed to reach the minimum sample size that would allow us to detect a small effect size ($w = 0.17$) in each pairwise comparison using the chi-square test and assuming $\alpha = .05/3$ and $1 - \beta = .80$ (Cohen, 1988). This meant 182 participants in each condition, 364 participants per comparison and 546 in total. The recruited sample size included a possible attrition rate (~10%) and reached 600 participants. One participant was excluded because of spending <30 s on the questionnaire. The analytical sample size was thus $N = 599$. The participants were recruited from an online UK panel (Prolific). To be eligible, they had to be at least 18 years old, reside in the UK and have an approval rating of >90%. A high approval rating should ensure a high quality of responses (Peer, Vosgerau, & Acquisti, 2014). The participants were paid £0.30 for completing a 3-min questionnaire.

Participants' ages ranged from 18 to 79 years ($M = 42.7$, $SD = 14.7$ years) and 49.6% were women, 49.4% were men and 1.0% were of other gender identities. The levels of the participants' education were relatively heterogeneous: 1.5% did not complete their high school education, 36.2% completed high school education, 46.6% completed a college degree, 13.4% completed a master's degree and 2.3% completed a PhD or other professional degree. The sample consisted of managers and working professionals (27.7%), unemployed people, including students and homemakers (15.5%); workers in sales and offices (10.7%), retired (9.5%), service workers (6.3%), government workers (6.0%) and some other, less common, occupations.

We used a simple between-subjects design with three conditions: single-event probability ($n = 205$), natural frequencies using standard wording ($n = 198$) and natural frequencies using improved wording ($n = 196$). Participants estimated the probability or frequency of having insulin-dependent diabetes given a positive genetic test. They were allocated to conditions randomly. In all experiments, the random allocation of the participants was done by the Qualtrics built-in randomiser, which operates automatically using the Mersenne Twister algorithm (Matsumoto & Nishimura, 1998).

2.2. Materials and procedure

After providing informed consent, participants were asked to calculate the posterior probability of having insulin-dependent diabetes given the positive results of genetic testing for the condition. We adopted the exact wording of the problem used in prior research for the single-event probability and the natural frequency with standard wording conditions (Galesic et al., 2009; Pighin et al., 2016). The numerical information was slightly changed to facilitate the mental calculation (we kept the original numbers in Experiments 2 and 3 for comparability). Specifically, the participants in the single-event probability condition were presented with probabilities expressed in percentages. (E.g., "The probability that a person has insulin-dependent diabetes is 1%.") They estimated the conditional probability that a person had insulin-dependent diabetes if they had a positive genetic test. (E.g., "Please estimate the probability that a person has insulin-dependent diabetes if he or she has a positive genetic test.")

The participants in the two natural frequency conditions were presented with frequencies with a natural sampling structure. (E.g., "100 out of every 10,000 people have insulin-dependent diabetes.") They estimated how many people had insulin-dependent diabetes out of those who got a positive genetic test (in the form of "__ out of __"). In the improved wording, however, we changed the description of the statistical information and the question (see Table 1). For example, the false-negative test results information was presented in the standard wording: "If a person does not have insulin-dependent diabetes, it is still possible that he or she will have a positive result on the genetic test. More precisely, 4,950 out of every 9,900 such people will have a positive result on the genetic test." In the improved wording condition, the same information was presented differently: "In addition to those 95 people, 4,950 people will also have a positive result on the genetic test out of the remaining 9,900 people who do not have insulin-dependent diabetes." Thus, the improved wording was designed to clearly associate the numbers with the reference categories and to cue the proper mathematical operation. (E.g., "In addition to those 95 people …") The question was presented in the standard wording as follows: "Here is a new representative sample of people who got a positive result on the genetic test. Please estimate how many of these people actually have insulin-dependent diabetes." It was presented in the improved wording

as: "Please estimate how many people actually have insulin-dependent diabetes out of the people who got a positive result on the genetic test." Thus, the reworded question was aligned with the order of information in the response format (X out of Y). Please see the Supplementary Materials for the exact wording of the problem in all the conditions. Participants then answered socio-demographic questions concerning their age, gender, level of education and occupation and were debriefed.

We used an accuracy criterion to categorise answers as correct or incorrect Bayesian reasoning (Vallée-Tourangeau, Sirota, Juanchich, & Vallée-Tourangeau, 2015). Specifically, we implemented a strict coding criterion whereby we coded only the exact numbers corresponding to the normatively correct answers as Bayesian answers. For the probability answers, we accepted answers with zero, one or more decimal places (e.g., 1.883, 1.9 or 2) as correct answers. This was because we did not request an exact number of decimal places in the task. A strictly accurate system of coding was chosen rather than approximately accurate answers since different non-Bayesian reasoning strategies in these problems have led to approximately correct answers (Galesic et al., 2009; Pighin et al., 2016). As a secondary measure, we calculated the absolute deviation of the estimate from the correct answer since manipulation could trigger a diverse array of non-Bayesian strategies resulting in overall less or more accurate estimates.

To test the effect of the manipulation on Bayesian reasoning, we used a series of chi-square tests with Yates' continuity correction (for the Bayesian/non-Bayesian answer variable) and Mann-Whitney *U* tests (for the absolute deviation from the objective value variable). To be able to quantify support for the models assumed by both the null and alternative hypotheses, we carried out equivalent Bayes factor analyses: a BF contingency table using an independent multinomial sampling plan with default prior concentration parameter, $a = 1$, and a Bayesian Mann-Whitney *U* test with a Cauchy scale of 0.707 and MCMC sampling with 5 chains and 10,000 iterations (set seeds 1). All analyses were conducted in R (version 4.1.1) except for the Bayesian Mann-Whitney *U* tests, which were conducted in JASP (version 0.13.1.0) (Morey & Rouder, 2015).

2.3. Results and discussion

Across the conditions, only a few participants calculated the Bayesian screening problem correctly. However, we observed considerable differences between the conditions (Table 2). We excluded the value of one participant in Experiment 1 from the absolute deviation variable for providing impossible values (the numerator was larger than the denominator) that was above 100%. We still coded the participant's answer as an incorrect value for the main variable of interest. This participant was from the improved wording condition. Only a handful of participants provided Bayesian answers in the probability and natural frequency conditions using the standard wording; their performance was close to zero. On the other hand, participants provided more Bayesian answers in the natural frequency condition using the improved wording (see Table 2). The omnibus test confirmed the existence of statistically significant differences between the three conditions, $\chi^2(2) = 22.91$, $p < .001$, Cramer's $V = 0.20$.

To test our hypotheses, we conducted three pairwise comparisons (we used the Bonferroni correction and adjusted $\alpha = .05/3 = .017$). We found no significant difference between the single-event probability condition and the natural frequency condition with standard wording, $\chi^2(1) = 1.68$, $p = .195$, Cramer's $V = 0.06$, with $BF_{01} = 7.6$ to 1, favouring no association between the statistical format and the number of correct responses. The improved wording of the natural frequency format yielded significantly more Bayesian answers than the single-event probability one, $\chi^2(1) = 7.98$, $p = .005$, Cramer's $V = 0.14$, yielding $BF_{10} = 6.4$ to 1 in favour of the association between the two types of problem and the number of correct responses. The improved (vs standard) wording also facilitated Bayesian reasoning in the problems featuring natural frequencies, $\chi^2(1) = 16.22$, $p < .001$, Cramer's $V = 0.20$, yielding $BF_{10} = 11.3*10^2$ to 1 in favour of the association between the two different types of wording and the number of correct responses.

Table 2
The effect of statistical format and problem-wording on Bayesian reasoning.

Experiments/Conditions	Bayesian answers strict criteria % (x/n)	Absolute deviation in probability estimate *Mdn* (*IQR*)
Experiment 1		
Single-event probability	3.4% (7/205)	53.1 (47.0)
NF – standard wording	1.0% (2/198)	1.2 (47.2)
NF – improved wording	11.2% (22/196)	0.9 (47.7)
Experiment 2		
NF – standard wording	1.9% (5/261)	3.9 (48.6)
NF – improved wording	9.6% (25/260)	0.5 (49.1)
NF – the same sample	3.1% (8/261)	49.0 (94.3)
NF – causal explanation	3.5% (9/256)	1.04 (48.6)
Experiment 3		
Single-event probability	2.7% (7/260)	49.3 (47.6)
NF – standard wording	1.5% (4/262)	39.0 (48.8)
NF – combined problem presentation	12.3% (32/260)	0.5 (48.6)
Experiment 4		
NF – standard wording	1.5% (4/262)	3.5 (30.5)
NF – improved wording	27.6% (74/268)	1.4 (13.0)
Experiment 5		
NF – standard wording	1.9% (4/209)	1.3 (5.2)
NF – improved wording	21.1% (44/209)	1.3 (6.6)
NF – reworded question	6.0% (13/217)	6.3 (77.2)
NF – reworded information's description	11.2% (24/214)	1.3 (0.1)
NF – arithmetic problem	86.7% (736/849)	–
Experiment 6		
NF – standard wording	4.0% (7/174)	20.8 (4.4)
NF – improved wording	22.7% (39/172)	19.4 (25.8)
Normalised F – improved wording	4.3% (7/164)	18.3 (15.1)
Experiment 7		
Single-event probability	0.5% (1/208)	47.2 (23.4)
NF – standard wording	4.0% (8/199)	27.8 (1.9)
NF – improved wording	13.4% (26/194)	26.0 (16.6)

NF – natural frequencies, (x/n) = the number of Bayesian answers out of all responses, *Mdn* = median, *IQR* = interquartile range. The combined improvement in Experiment 3 comprises improved wording, the same sample and a causal explanation of the false-positive values.

Could the absence of the natural frequency effect be explained by applying a strict coding rule? It might appear so when looking at Fig. 1. However, a closer look at the answers adjacent to the Bayesian ones reveals that these are not miscalculations but mostly base-rate-only answers. Consider, for example, the responses between 1% and 2% in Experiment 1, where the prior probability is 100/10000 and the posterior probability is 95/5045. Only two out of 84 such answers were coded as Bayesian answers in the standard wording natural frequencies condition. Of the remaining 82 incorrect values, 70 are clearly base-rate-only answers (i.e., mostly 100 out of 10000; but also some 10 out of 1000; 1 out of 100 ratios); 10 values seem to be unclear strategies that can't be considered miscalculations (e.g., 1 out of 50; 1 out of 95; 2 out of 100; 200 out of 15,000; 49.5 out of 4950) and 2 answers might potentially be considered miscalculations/rounding errors (95 out of 5050). Even if this were the case, this is a negligible number of false negative answers. Thus, a lenient coding rule would not offer a good trade-off: the cost of introducing a large number of false positives would severely outweigh the benefit of including a small number of false negatives.

Participants' estimates varied in terms of median absolute deviation from the correct value, *K-W*(2) = 86.14, $p < .001$ (Table 2). (We excluded the value of one participant from the absolute deviation variable for providing implausible value where the numerator was larger than the denominator. We still coded the participant's answer as an incorrect value for the main variable of interest.) The estimates deviated

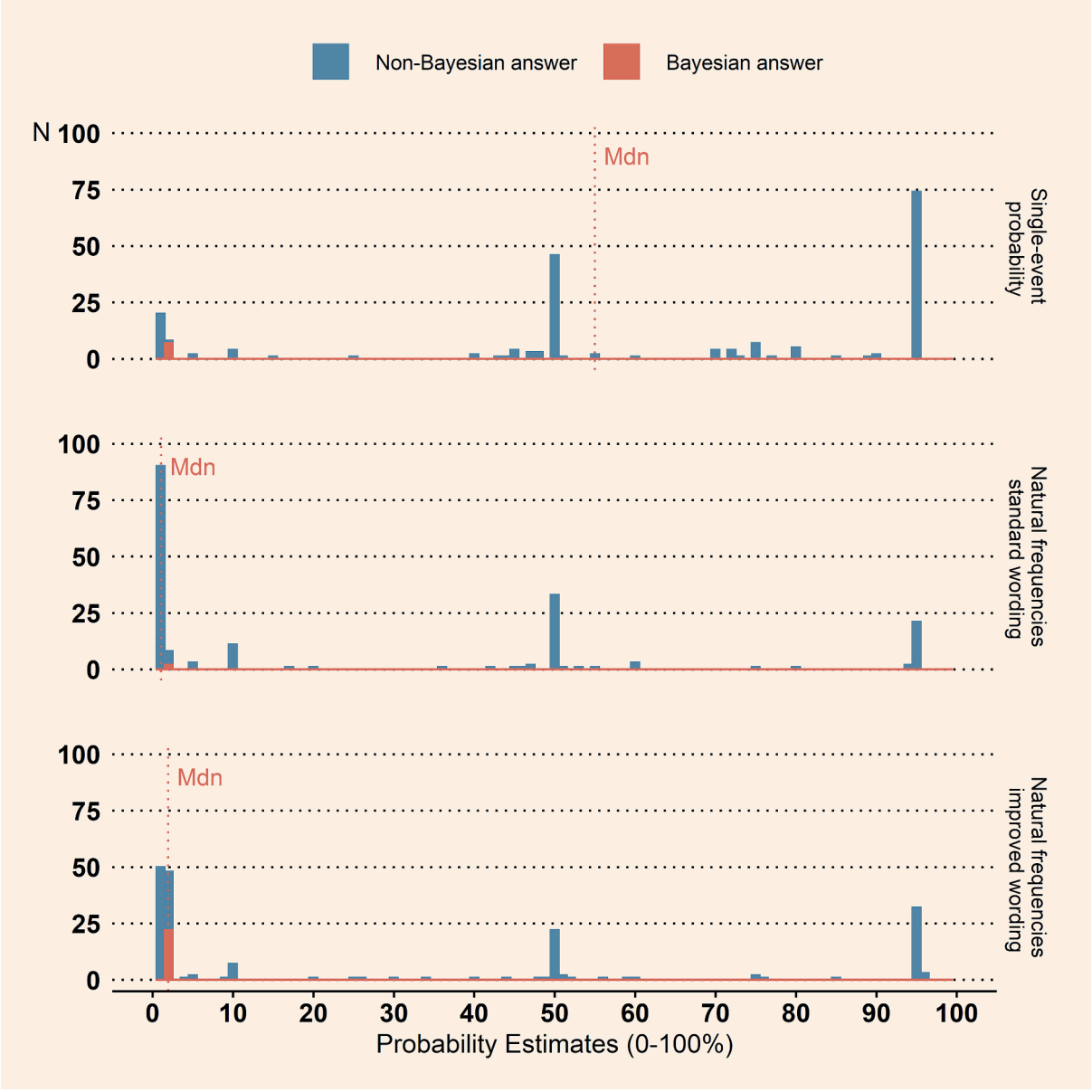

Fig. 1. Effect of statistical format and wording on Bayesian reasoning.
The dotted vertical line represents the median probability estimate (*Mdn*) within each condition.

less from the correct value in the natural frequency conditions than with the single-event probabilities, likely due to the employment of different sets of non-Bayesian strategies as visible in the histograms of the answers per condition (Fig. 1). Three pairwise comparisons confirmed the expectations (using, as previously, an adjusted $\alpha = .017$). We found a rank-sum difference between the single-event probability and the natural frequencies with the standard wording, $W = 29{,}797$, $p < .001$, yielding $BF_{10} = 13.0{*}10^5$ to 1 in favour of the difference in absolute deviation according to the format. We also found a difference between the single-event probability and the natural frequencies with the improved wording, $W = 28{,}716$, $p < .001$, yielding $BF_{10} = 37.1{*}10^4$ to 1 in favour of the difference in absolute deviation according to the format. Finally, the improved and standard wording of the natural frequencies did not differ significantly in the median absolute deviation from the correct value, $W = 20{,}089$, $p = .483$, $BF_{01} = 4.4$ to 1 in favour of no difference in absolute deviation according to the wording.

To summarise, the natural frequencies did not facilitate Bayesian reasoning, relative to the single-event probabilities, when we used the problem with the standard wording. Similar findings were reported in the general adult population before with medical screening tasks (Pighin et al., 2016). However, once we improved the wording of the problem, using natural frequencies was beneficial and the normative performance increased. Thus, the wording appeared to be the critical component of the facilitatory effect of the natural frequencies in this problem.

3. Experiment 2

In Experiment 1, we found a significant effect of wording on Bayesian reasoning when used with the medical screening test problem featuring natural frequencies. However, it is unclear how strong the effect is relative to other problem representation changes known to improve Bayesian reasoning. In Experiment 2, therefore, we aimed to compare the effect of problem-wording on Bayesian reasoning with the standard wording but also with the same sample type effect (Johnson & Tubau, 2017) and the impact of providing a causal explanation (Krynski & Tenenbaum, 2007). Specifically, to manipulate the sample type,

participants were asked to solve the problem using either a new or the same sample as described in the problem (Johnson & Tubau, 2017). To manipulate provision of causal explanation, participants read (or not) an explicit alternative cause for false positive tests, namely, that the tests are detecting harmless gene variants that do not cause diabetes (Krynski & Tenenbaum, 2007).

Based on the previous literature, we hypothesised that the improved wording would yield more Bayesian answers, relative to the standard wording problem presentation, and a larger or similar level of Bayesian answers relative to the same sample type effect and causal explanation. (The similar level of performance was assumed in the case these two methods were effective relative to the standard wording in the screening context.)

3.1. Method

3.1.1. Participants and design

We aimed to reach the minimum sample size that would allow us to detect a small effect size ($w = 0.15$) in each pairwise comparison using the chi-square test and assuming $\alpha = .017$ (.05/3) and $1 - \beta = .80$ (Cohen, 1988). This was a conservative effect size estimate given the effect found in the prior study ($w = 0.20$). This meant a target of 233 participants per condition to reach 466 participants per comparison, adding up to an overall sample of 932 participants across four conditions. The recruited sample size included a possible attrition rate and reached 1038 participants (~260 participants per condition). The participants were recruited from an online UK panel (Prolific). To be eligible, they had to be at least 18 years old, reside in the UK and have an approval rating of >90%. A high approval rating should ensure a high quality of responses (Peer et al., 2014). The participants were paid £1 for completing a 12-min questionnaire that featured unrelated tasks reported elsewhere (Sirota, Thorpe, & Juanchich, 2022).

Participants' ages ranged from 18 to 84 years ($M = 36.8$, $SD = 11.8$ years), and 70.9% were women, 28.8% were men and 0.3% were of other gender identities. The education levels achieved by the participants were relatively heterogeneous: 0.9% did not complete their high school education, 41.7% completed high school education, 43.2% completed an undergraduate degree, 11.1% completed a master's degree and 3.2% completed a PhD or other professional degree. The sample consisted of managers and working professionals (26.4%), unemployed people, students and homemakers (20.2%); workers in sales and offices (12.0%), services (7.9%), government (4.8%), retired (4.8%) and some other, less common, occupations.

Participants estimated the frequency of having insulin-dependent diabetes given a positive genetic test in a simple between-subjects design with four conditions: (i) natural frequencies with standard wording ($n = 261$), (ii) natural frequencies with improved wording ($n = 260$), (iii) natural frequencies with a question using the same sample ($n = 261$) and (iv) natural frequencies with a causal explanation of false-positives ($n = 256$). They were allocated to conditions randomly.

3.1.2. Materials and procedure

After providing informed consent and completing unrelated tasks, participants were asked to interpret the positive results of genetic testing for insulin-dependent diabetes in one of the four conditions. In all the conditions, the problem featured the same statistical information using the natural frequency format. The wording of the problem differed across the conditions. (See the exact wording of each problem per condition in the Supplementary Materials.) We used the same problems as in Experiment 1 for the standard and improved wording conditions. In the same sample type condition, the standard wording was used but the question was changed to ease the relational mapping demand by applying the subset structure to the same sample: "Please estimate how many of the people who got a positive result on the genetic test actually have insulin-dependent diabetes," (Johnson & Tubau, 2017). In the causal explanation condition, the standard wording was accompanied by a causal explanation of false-positives: "This is because some gene changes usually associated with the diagnosis might just be harmless gene variants that do not cause diabetes," (Krynski & Tenenbaum, 2007). Participants then answered questions concerning their age, gender, level of education and occupation.

The same coding strategy and statistical analyses were performed as in Experiment 1.

3.2. Results and discussion

Our participants provided only a few Bayesian answers when presented with the standard wording of the natural frequency version of the problem. They provided only slightly more correct answers when the problem representation was enhanced using the same type of sample between the problem and the question and a causal explanation (Table 1). However, we observed a noticeable improvement in the proportion of Bayesian answers in the improved wording condition. We found a statistically significant difference between the four conditions, $\chi^2(3) = 21.59$, $p < .001$, Cramer's $V = 0.14$. We used pairwise comparisons to test our hypotheses (with a Bonferroni correction $\alpha = .05/3 = .017$). We found a significantly higher proportion of Bayesian answers with the improved wording compared to the standard wording, $\chi^2(1) = 12.85$, $p < .001$, Cramer's $V = 0.16$, yielding $BF_{10} = 87.3$ to 1 in favour of the association between the two types of wording and the number of correct responses. We further found that the improved wording was more effective than the same type sample wording, $\chi^2(1) = 8.35$, $p = .004$, Cramer's $V = 0.13$, $BF_{10} = 6.4$ and the causal explanation, $\chi^2(1) = 6.84$, $p = .009$, Cramer's $V = 0.12$, $BF_{10} = 2.8$. In addition to the main wording-effect hypotheses, we found no support for the effectivity of the same sample type and causal explanation relative to the standard wording, $\chi^2(1) = 0.32$, $p = .574$, Cramer's $V = 0.02$, $BF_{01} = 20.7$; $\chi^2(1) = 0.72$, $p = .396$, Cramer's $V = 0.04$, $BF_{01} = 15.2$, respectively. The lack of effect can explain the fact that the improved wording was more effective (rather than similarly effective) than the same type of sample effect and causal explanation effect.

In terms of absolute deviation from the correct numerical answer, the responses of the participants in the improved wording condition were the closest to the correct response, followed by the causal explanation and standard wording conditions. However, the participants in the same type of sample condition produced answers that differed the most from the correct answers (Table 2, Fig. 2). Such deviations were driven by different subsets of non-Bayesian strategies employed across the conditions. For instance, the participants in the improved wording condition generated more Bayesian answers (which was a low probability) and fewer non-Bayesian answers yielding medium and high probabilities (see Fig. 2). We found a statistically significant omnibus difference between the conditions, $K\text{-}W(3) = 48.24$, $p < .001$. We used pairwise comparisons to test our hypotheses (with an adjusted $\alpha = .017$). The responses deviated less from the correct answer in the natural frequencies with the improved wording than with the standard wording, $W = 37{,}000$, $p = .022$, but this was not a statistically significant difference using the alpha corrected for multiple comparisons. We found only $BF_{10} = 1.5$ to 1 in favour of the difference according to the wording in the absolute deviation from the normatively correct answer. The improved wording led to a smaller deviation from the correct answers compared to the same type sample condition, $W = 22{,}838$, $p < .001$, $BF_{10} = 16.3*10^3$ and the causal explanation condition, $W = 28{,}541$, $p = .009$, $BF_{10} = 4.2$.

Thus, this experiment replicated the effect of improved wording on Bayesian reasoning. The improved wording was more effective than the other two presentation enhancement methods tested: the same sample type presented in the problem and the question and the causal explanation of false positives.

4. Experiment 3

In Experiment 2, we replicated the effect of the improved wording,

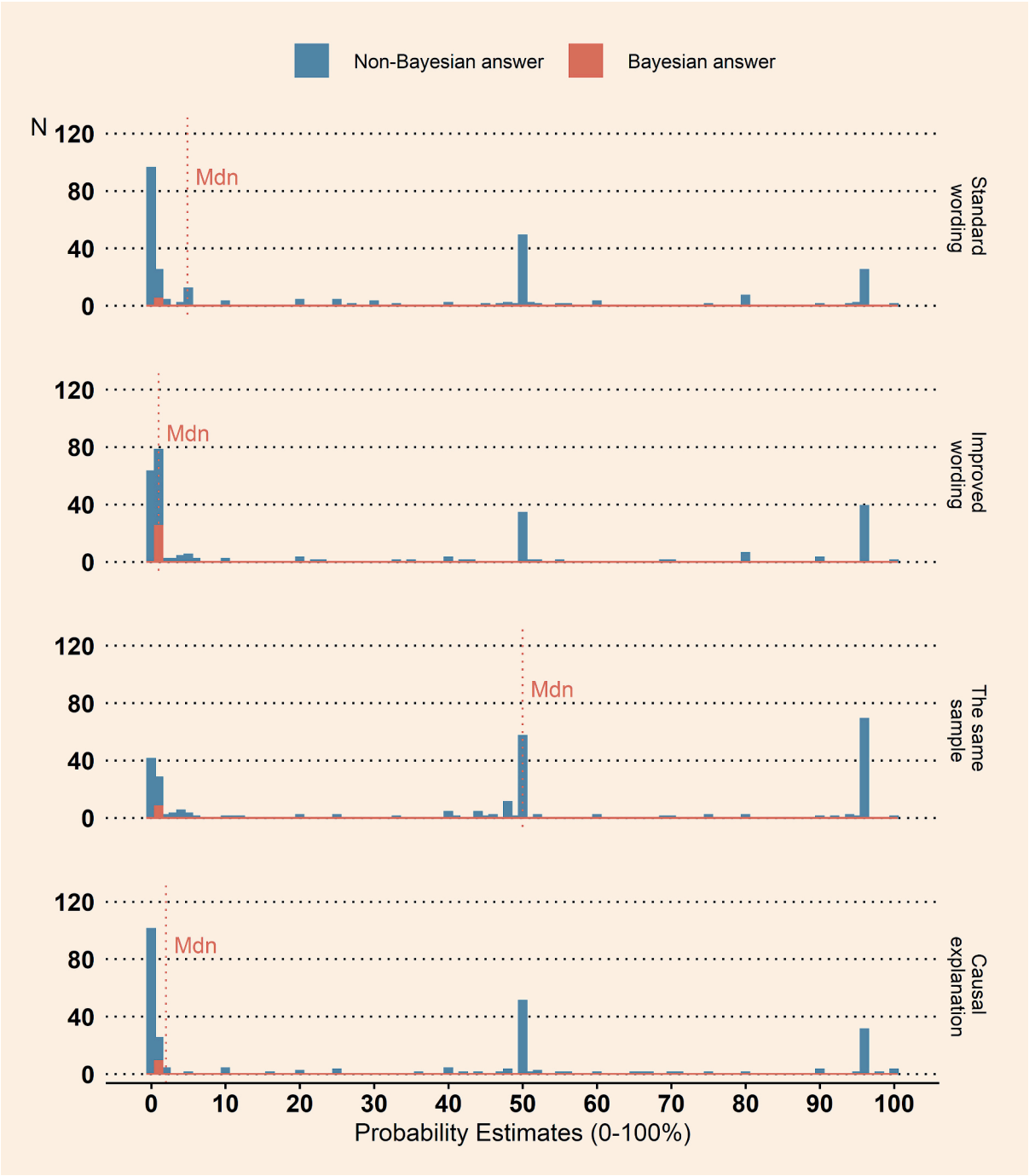

Fig. 2. Effect of wording and problem representation changes on Bayesian reasoning with natural frequencies.
The dotted vertical line represents the median probability estimate (Mdn) within each condition.

relative to the standard wording, and we found that it boosted Bayesian reasoning more effectively than other problem-enhancement methods. In Experiment 3, we extended these findings by testing whether the facilitatory effects of these enhancement methods could be cumulative. Therefore, we tested a complex problem intervention featuring improved wording, the same type of sample and a causal explanation. The cumulative effect would have practical implications for effective risk communication in medical practice. Another aim of this experiment was to retest the surprising lack of the facilitatory effect of the natural frequencies using the standard wording, relative to the single-event probability format, identified in Experiment 1. Based on the prior findings, we hypothesised that the improved problem presentation would facilitate Bayesian reasoning relative to the natural frequencies with the standard wording and the single-event probability format. We also predicted that the improved problem representation designed to enhance problem comprehension would have an additional facilitatory effect for problems featuring natural frequencies.

4.1. Method

4.1.1. Participants and design

This experiment was run in parallel with Experiment 2; the same power considerations, recruitment strategy, reimbursement and eligibility criteria were followed as in Experiment 2. The actual sample size

reached 782 participants (~260 per condition for three conditions). Participants' ages ranged from 18 to 73 years (M = 36.3, SD = 11.2 years) and 69.6% were women, 30.1% were men and 0.4% were of other gender identity. The levels of the participants' education were relatively heterogeneous: 1.8% did not complete their high school education, 40.8% completed high school education, 43.0% completed a college degree, 12.0% completed a master's degree and 2.4% completed a PhD or other professional degree. The sample consisted of managers and working professionals (24.4%), unemployed people, including students and homemakers (25.4%); workers in sales and offices (11.0%), services (6.8%), government (5.4%) and some other, less common, occupations.

We used a simple, between-subjects design with three conditions: single-event probability (n = 260), natural frequencies using standard wording (n = 262) and natural frequencies using an improved problem presentation (n = 260). Participants estimated the probability or frequency of having insulin-dependent diabetes given a positive genetic test. They were allocated to conditions randomly.

4.1.2. Materials and procedure

After providing informed consent and completing unrelated tasks, participants were asked to interpret the positive results of genetic testing for insulin-dependent diabetes. Depending on the format condition, participants either estimated the conditional probability that a person had insulin-dependent diabetes, given that they received a positive genetic test, or how many people had insulin-dependent diabetes out of those who got a positive genetic test. The problems in the single-event probability condition and the natural frequency using the standard wording condition were identical to those in Experiment 1 but featured the numbers appearing in Experiment 2. The problem in the improved problem representation condition was presented using the same improved wording as Experiment 2, but it also featured a causal explanation of false positives and the question presenting the same sample as in the problem rather than a new sample (see Table 3). Participants then answered socio-demographic questions concerning their age, gender, level of education and occupation and were debriefed.

Table 3
The standard wording and combined improvement in problem presentation (improved wording, same sample type and causal explanation) of the diabetes screening problem.

Standard Wording	Combined Improvement Representation
To determine whether a person is at risk of insulin-dependent diabetes, doctors sometimes conduct genetic testing. If a person tests positive for a certain gene, he or she might have insulin-dependent diabetes. Here is some information about that genetic test.	To determine whether a person is at risk of insulin-dependent diabetes, doctors sometimes conduct genetic testing. If a person tests positive for a certain gene, he or she might have insulin-dependent diabetes. Here is some information about that genetic test.
50 out of every 10,000 people have insulin-dependent diabetes.	50 out of every 10,000 people have insulin-dependent diabetes.
If a person has insulin-dependent diabetes, it is not sure that he or she will have a positive result on the genetic test. More precisely, 48 of every 50 of such people will have a positive result on the genetic test.	**48 people will have a positive result on the genetic test out of the 50 people who have insulin-dependent diabetes.**
If a person does not have insulin-dependent diabetes, it is still possible that he or she will have a positive result on the genetic test. More precisely, 4975 out of every 9950 such people will have a positive result on the genetic test.	**In addition to those 48 people, 4975 people will also have a positive result on the genetic test out of the remaining 9950 people who do not have insulin-dependent diabetes.** This is because some gene changes usually associated with the diagnosis might just be harmless gene variants that do not cause diabetes.
Here is a new representative sample of people who got a positive result on the genetic test. Please estimate how many of these people actually have insulin-dependent diabetes.	***Please estimate how many people actually have insulin-dependent diabetes out of the people who got a positive result on the genetic test.***

The sentence "This is because..." represents the text clarifying the causal model. The *text in italics* indicates the same sample type (i.e., the question refers to the people described previously, not to a new representative sample of people). The **text in bold** indicates the improved wording aiming to enhance text comprehension (i.e., the close association of the numbers with the reference categories, clear cueing mathematical operations, simplifying the sentences by removing conditional sentences and matching the order in the response question).

4.2. Results and discussion

Participants provided only a few Bayesian answers when presented with a single-event probability version, and a natural frequency version of the problem using standard wording. The performance in both conditions was close to zero (see Table 2). The improved problem presentation of the diabetes problem had substantially (+ 10–11%) improved Bayesian answers (Table 2). We found a statistically significant difference between the three conditions, $\chi^2(2) = 35.09$, $p < .001$, Cramer's $V = 0.21$, which we followed up with three pairwise comparisons to test our hypotheses. (We used the Bonferroni correction and adjusted $\alpha = .05/3 = .017$.) We found no significant difference between the single-event probability version and the natural frequency version with the standard wording, $\chi^2(1) = 0.39$, $p = .534$, Cramer's $V = 0.03$, with $BF_{01} = 20.9$ to 1, favouring no association between the statistical format and the number of correct responses. The natural frequency version with the improved problem presentation—the complex manipulation consisting of improved wording, the same sample type and a causal explanation—yielded significantly more Bayesian answers than the single-event probability one, $\chi^2(1) = 15.97$, $p < .001$, Cramer's $V = 0.18$, yielding $BF_{10} = 50.3*10^1$ and the standard wording problem featuring a natural frequency, $\chi^2(1) = 21.97$, $p < .001$, Cramer's $V = 0.21$, yielding $BF_{10} = 21.5*10^3$. We also compared the improved problem representation of Experiment 3 with the improved wording condition of Experiment 2, which was run in parallel and thus enabled us to draw causal inferences. We found no significant differences between the conditions, $\chi^2(1) = 0.71$, $p = .400$, Cramer's $V = 0.04$, yielding $BF_{01} = 9.1$, favouring no association between the two conditions and the number of correct responses. This means that the effect of the improved problem presentation was completely driven by the improved wording.

Furthermore, participants' deviation from the correct value varied across the conditions (Fig. 3). Their estimates in the improved problem presentation condition were substantially closer to the normative answer than the natural frequencies with the standard problem representation of natural frequencies format or the single-event probability version of the task (Table 2). The omnibus difference between the three conditions was statistically significant, $K\text{-}W(2) = 84.82$, $p < .001$. Three pairwise comparisons identified the improved presentation of the natural frequencies as the main driving force of this difference (using, as previously, an adjusted $\alpha = .017$). We found a rank-sum difference between the single-event probability and the natural frequencies with the standard wording, $W = 43{,}955$, $p < .001$, yielding $BF_{10} = 68.8*10^2$ to 1 in favour of the difference in absolute deviation according to the format. We also found a difference between the single-event probability and the natural frequencies with the improved wording, $W = 48{,}056$, $p < .001$, yielding $BF_{10} = 65.0*10^5$ to 1 in favour of the difference in absolute deviation according to the format. Finally, the natural frequencies with the improved problem presentation were significantly more aligned with the correct response than the natural frequencies with the standard wording, $W = 39{,}044$, $p < .001$, $BF_{10} = 26.3$ to 1 in favour of the difference in absolute deviation according to the format.

To summarise, we found mixed support for the hypothesis that natural frequencies facilitate Bayesian reasoning. Natural frequencies did not facilitate the normative performance using the standard wording, relative to the single-event probabilities (similar to Pighin et al., 2016); they improved the performance only when the problem presentation was substantially enhanced. As predicted, the hypothesis suggesting the

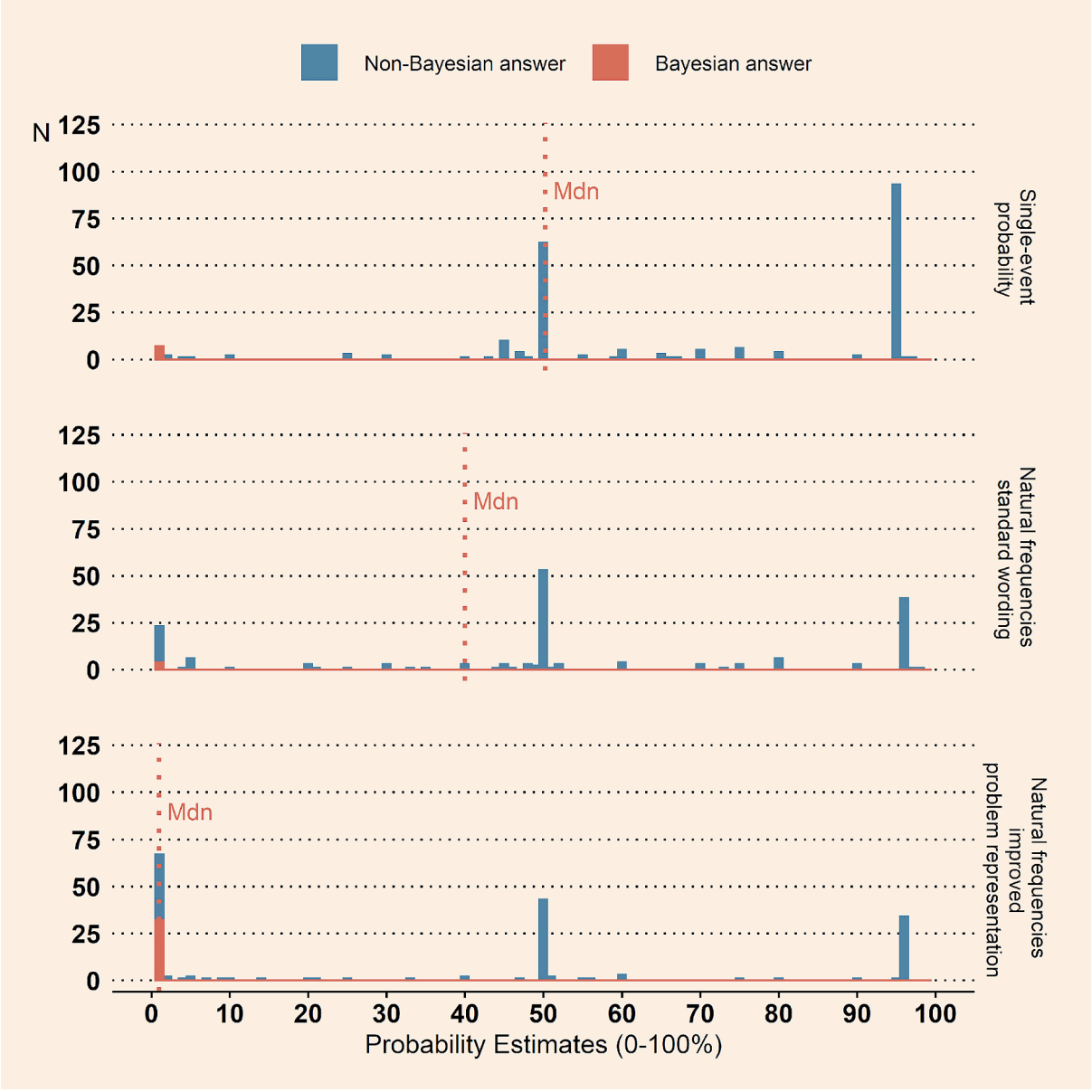

Fig. 3. Effect of statistical information and problem representation on Bayesian reasoning.
The dotted vertical line represents the median probability estimate (*Mdn*) within each condition.

problem presentation effect was supported. The combined improved problem presentation yielded better normative performance and a much smaller absolute deviation than the standard wording of the problem featuring single-event probabilities and natural frequencies.

5. Experiment 4

In Experiment 4, we tested whether the wording effect generalised to other Bayesian textbook medical problems. For that purpose, we used a Trisomy 21 textbook problem (Galesic et al., 2009; Pighin et al., 2016) and applied the changes to the wording following the same principles as described above in Experiment 1 (i.e., the verbal description of the statistical information and question). In this preregistered experiment, we hypothesised that the improved wording of the natural frequency version of the Trisomy 21 problem would increase the proportion of Bayesian answers and yield estimates with lower absolute deviations from the correct response compared with the standard wording of the natural frequency version of the Trisomy 21 problem.

5.1. Method

5.1.1. Participants and design

We aimed to recruit at least 528 participants to detect the smallest effect of the wording observed in the previous studies ($w = 0.157$), while using the chi-squared independence test and assuming $\alpha = .05$, $1 - \beta = .95$, and a two-sided test (Cohen, 1988). Therefore, to compare the two conditions, we recruited 530 participants from the online panel Prolific to complete an online questionnaire. (The two additional participants were due to the recruitment method.) Only participants who were at least 18 years old and UK nationals currently residing in the UK were eligible to participate. The participants were paid £0.50 for completing a 5-min questionnaire.

Participants' ages ranged from 18 to 71 years ($M = 35.9$, $SD = 12.5$ years), and 71.1% were women, 28.7% were men, and 0.2% were of other gender identities. The levels of education achieved by the participants were relatively heterogeneous: 0.2% did not complete their high school education, 43.8% completed high school education, 40.9%

completed an undergraduate degree, 11.5% completed a master's degree and 3.6% completed a PhD or other professional degree. The sample consisted of unemployed people, students and homemakers (27.5%); managers and working professionals (22.5%), workers in sales and offices (12.8%), government (6.6%), services (5.8%) and retired (4.2%) or some other occupation category.

In a simple between-subjects design, participants solved the Trisomy 21 medical screening problem featuring natural frequencies, using either the standard ($n = 262$) or improved wording ($n = 268$). They were allocated to conditions randomly.

5.1.2. Materials and procedure

After providing informed consent, participants read and solved the Trisomy 21 problem, describing a neck-fold prenatal test assessing the risk of an unborn child having Down syndrome (Galesic et al., 2009). Participants then estimated the number of women having a child with Down syndrome out of those with a positive screening test. The standard wording of the problem was taken from prior research (Galesic et al., 2009). The improved wording of the problem followed the same principles as the "diabetes" problem, resulting in a change to the verbal description of the statistical data and the question format. (See Supplementary Material for the full wording.) Afterwards, participants answered questions concerning their age, gender, level of education and occupation.

As in the previous experiments, we used a strict criterion to code normatively correct answers corresponding to the Bayesian calculations (i.e., 12/(12 + 799)) and the absolute deviation from the correct answer, reflecting the degree of (in)accuracy of the answer. The same statistical analyses were performed as in Experiment 1.

5.2. Results and discussion

As predicted, participants provided many more Bayesian answers when presented with the improved wording compared with the standard wording of the natural frequency version of the task, which was close to zero (see Table 2). Following the preregistered analytical plan, this 26%

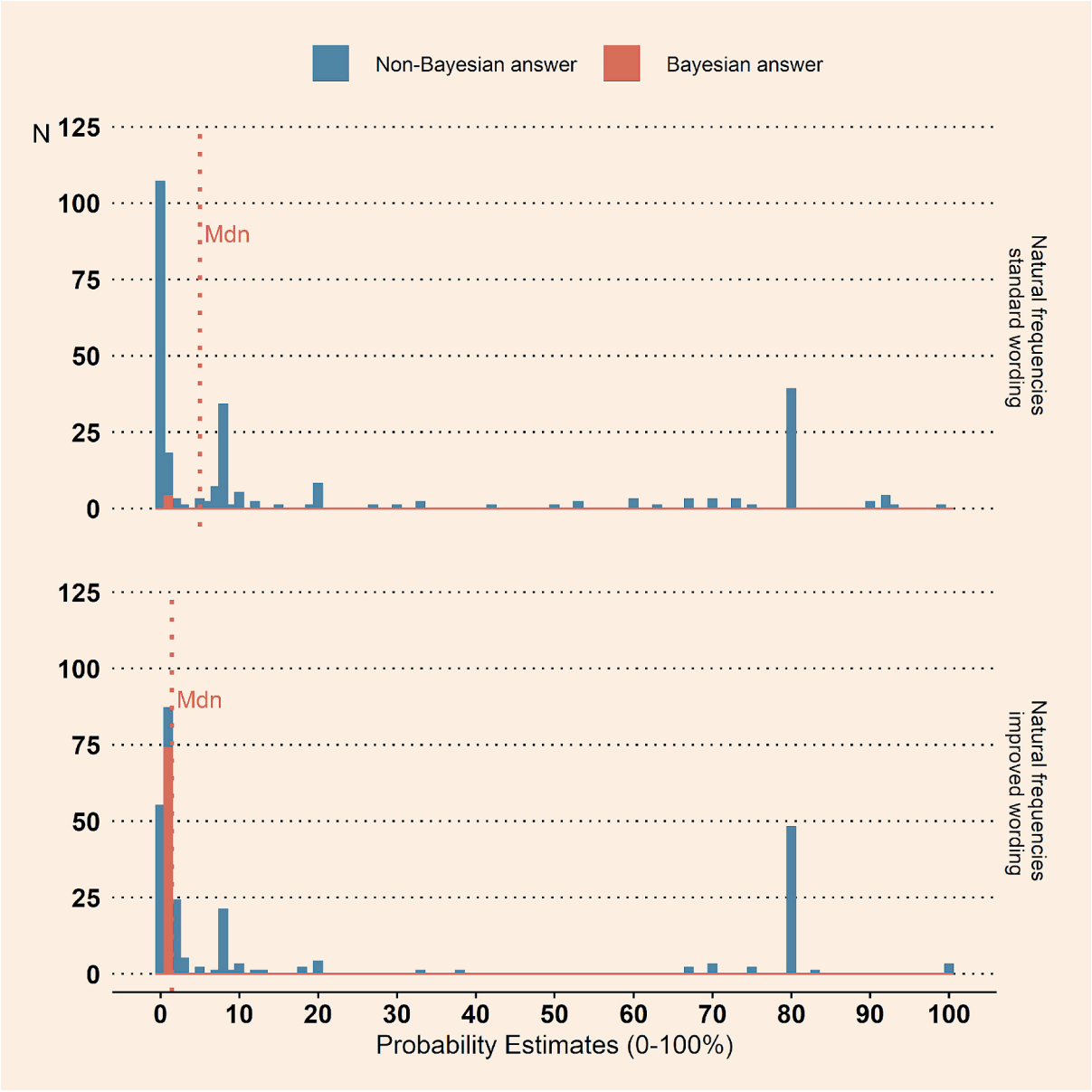

Fig. 4. Effect of problem-wording on Bayesian reasoning with natural frequencies.
The dotted vertical line represents the median probability estimate (Mdn) within each condition.

difference was statistically significant, $\chi^2(1) = 69.76$, $p < .001$, Cramer's $V = 0.36$, yielding $BF_{10} = 13*10^{16}$ to 1 in favour of the association between the different problem presentations and the number of correct responses. Similarly, participants deviated less from the Bayesian answer when presented with the improved wording compared to the standard wording (Table 2; Fig. 4). The effect of the wording was statistically significantly different, $W = 43{,}690$, $p < .001$, yielding $BF_{10} = 27*10^2$ to 1 in favour of the difference according to the wording in the absolute deviation from the normatively correct answer. This evidence supports the hypotheses of the facilitative effect of the improved wording relative to the standard wording.

6. Experiment 5

Across four experiments, we showed that the improved wording facilitated Bayesian reasoning. In this preregistered experiment, we aimed to test whether this was due to the wording changes in how the statistical information was described, the question that was asked, or both. We also aimed to parse the two challenges that people might face in solving word problems: understanding the task and computing the correct answer. To identify how much the mathematical computation prevents people from reaching the correct answer, we asked participants to complete a simple numerical task—computationally identical to the one presented in the Trisomy 21 problem, but presented as an arithmetic expression. Having the numeric task allowed us to disentangle the computational difficulties in solving the problem from the problem representational difficulties people might have when solving the Bayesian problem (Johnson & Tubau, 2015). We hypothesised that the improved wording—of the statistical information description, the question and both—would increase Bayesian reasoning relative to the standard wording. We expected that each component of the improved wording would contribute to this increase. We also hypothesised that participants would not perform as well in the word problem as in the numerical version of the problem, accounting for the role of verbal problem comprehension.

6.1. Method

6.1.1. Participants and design

We aimed to recruit at least 832 participants based on an a-priori power analysis. For each comparison, 416 (i.e., 2*208 per condition) participants were needed to detect half of the effect size found in Experiment 4 (i.e., $w = 0.36/2 = 0.18$) while using a chi-squared test of independence to test the hypotheses and assuming $\alpha = .017$, $1 - \beta = .90$, a two-sided test, and a small attrition rate (Cohen, 1988). We thought half of the effect size would be reasonable to expect for the components of the wording effect: the reworded question and the reworded description of the statistical information. Therefore, 849 participants were recruited from an online panel, Prolific, and completed an online questionnaire. (One additional participant was excluded as they completed the study too quickly, i.e. in 15 s.) The sample was balanced in terms of participants' sex. Only participants who were at least 18 years old, and UK nationals currently residing in the UK, with at least a 90% approval rate of previous studies were eligible to participate. The participants were paid £0.30 for completing a 3-min questionnaire.

Participants' ages ranged from 18 to 80 years ($M = 40.6$, $SD = 13.3$ years), and 49.7% were women, 50.3% were men. The levels of education achieved by the participants were relatively heterogeneous: 1.3% did not complete their high school education, 38.2% completed high school education, 43.1% completed an undergraduate degree, 13.4% completed a master's degree and 4.0% completed a PhD or other professional degree. The sample consisted of managers and working professionals (30.0%), unemployed people, students and homemakers (17.2%); workers in sales and offices (11.7%), services (8.7%), retired (7.5%), government (6.2%) or some other occupation category.

In a simple between-subjects design, participants solved a medical screening scenario presented in the natural frequency format using one of the four types of wording: standard wording ($n = 209$), improved wording ($n = 209$), reworded question ($n = 217$) and reworded information description ($n = 214$). They were allocated to conditions randomly.

6.1.2. Materials and procedure

After providing informed consent, participants read and solved the Trisomy 21 problem describing a neck-fold prenatal test as an indicator of a child with Down syndrome (Galesic et al., 2009). Participants then estimated the number of women having a child with Down syndrome out of those with a positive screening test. The standard and improved wording conditions were identical to those used in Experiment 4. In the reworded question-only condition, only the wording of the question was improved. In the reworded-information-description condition, the wording of the verbal description of the statistical information was improved. (See Supplementary Material for the exact wording of the problems.) All participants were then asked to solve the numerical task computationally equivalent to the Bayesian problem. E.g., "Please solve the following problem: 12 out of (12 + 799)". They were then asked to provide their answers in the textboxes (__ out of __). Afterwards, the participants answered questions concerning their age, gender, level of education and occupation.

As in the previous experiments, we used a strict criterion to code normatively correct answers corresponding to the Bayesian calculations, i.e. 12/(12 + 799) and the absolute deviation from the normatively correct answer as a secondary measure. The same statistical analyses were performed as in previous experiments to test the hypotheses. In addition, the McNemar test was used to test differences between the estimates derived from the word problem and the numerical task.

6.2. Results and discussion

Participants provided more Bayesian answers when presented with the improved wording of the question, the description of the statistical information and both, compared to the standard wording of the natural frequency version of the task, which was close to zero (see Table 2). The performance, however, varied substantially across the improved wording conditions. The statistically significant omnibus test of differences, $\chi^2(3) = 47.73$, $p < .001$, Cramer's $V = 0.24$, was followed up by pairwise comparisons to test our hypotheses (using the Bonferroni correction $\alpha = 0.05/3 = 0.017$). As predicted, the participants generated significantly more Bayesian answers in the problem with the improved wording of both the verbal description of the statistical information and the question, than with the standard wording, $\chi^2(1) = 35.80$, $p < .001$, Cramer's $V = 0.29$, yielding $BF_{10} = 11.2*10^7$ to 1 in favour of the association between the two different wordings and the number of correct responses. The participants generated slightly more Bayesian answers in the problem with the improved wording of the question than with the standard wording, but this was not statistically significant, $\chi^2(1) = 3.62$, $p = .057$, Cramer's $V = 0.09$, yielding $BF_{01} = 2.1$ to 1 in favour of no association between the different types of wording and the number of correct responses. Finally, the participants generated significantly more Bayesian answers in the problem with the improved wording of the statistical information description than with the standard wording, $\chi^2(1) = 13.33$, $p < .001$, Cramer's $V = 0.18$, yielding $BF_{10} = 15.4*10^1$ to 1 in favour of the association between the two different types of wording and the number of correct responses. Thus, we confirmed the hypotheses about the overall wording effect and the effect of the reworded description of the statistical information, but not the effect of the reworded question.

Participants showed less deviation from the Bayesian estimate across the conditions (Table 2Fig. 5). This likely occurred because the participants used non-Bayesian strategies yielding low probability estimates close to the Bayesian estimation in this specific word problem. For instance, the conservatism strategy, i.e., 15 out of 10,000, occurred 181

times, and its variation, 12 out of 10,000, occurred 53 times (Zhu & Gigerenzer, 2006). Despite the small variability, we found a statistically significant omnibus difference between the conditions, K-$W(3) = 48.39$, $p < .001$. Following the preregistered analytical plan, we used pairwise comparisons to test our hypotheses (with an adjusted $\alpha = 0.017$). We found insufficient support for the difference using our adjusted alpha between the standard and improved wording, $W = 24{,}616$, $p = .018$, yielding $BF_{10} = 3.3$ to 1 in favour of the difference according to the wording in the absolute deviation from the normatively correct answer. In contrast with our expectation, the standard wording led to less deviation than the problem with the reworded question, $W = 18{,}016$, $p < .001$, yielding $BF_{10} = 3.9$ to 1 in favour of the difference according to the wording in the absolute deviation from the normatively correct answer. This was because participants employed different non-Bayesian strategies. (See Fig. 5, for instance. The representative thinking strategy "12 out of 15" was more common in the reworded question condition.) Lastly, as predicted, the problem with the reworded description of the statistical information generated fewer answers deviating from the normative answer than the standard wording, $W = 18{,}016$, $p < .001$, yielding $BF_{10} = 19.7$ to 1 in favour of the difference according to the wording in the absolute deviation from the normatively correct answer. Thus, we confirmed only the hypotheses about the positive effect of the reworded description of the statical information, but not the positive effect of the improved wording and the reworded question.

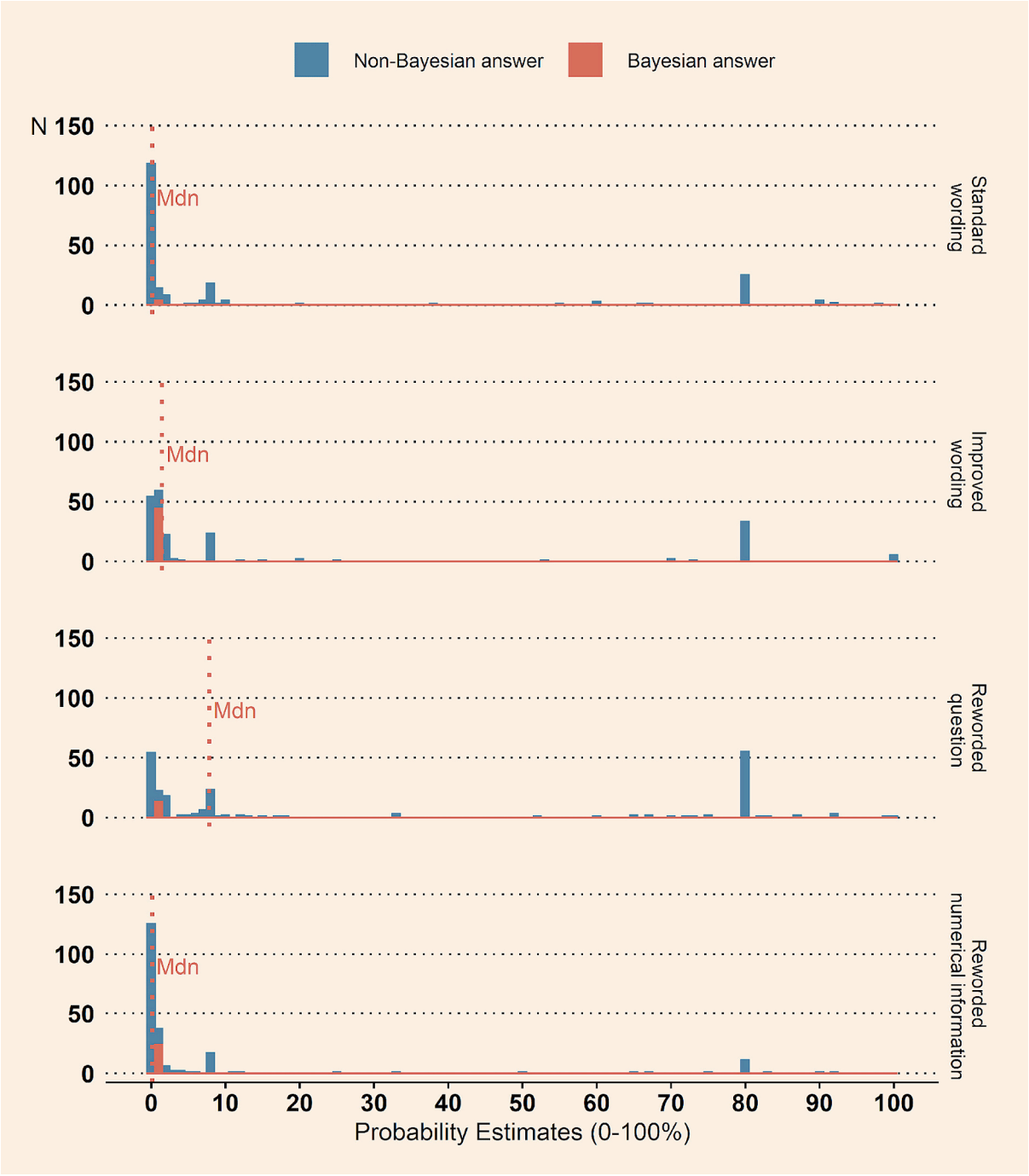

Fig. 5. Effect of problem-wording on Bayesian reasoning with natural frequencies.
The dotted vertical line represents the median probability estimate (*Mdn*) within each condition.

Finally, participants provided substantially and significantly more Bayesian answers in the problems presented arithmetically only (86.7%) compared to word problems where the same equation had to be extracted from the text (10.0%), McNemar's $\chi^2(1) = 647.01$, $p < .001$, Cohen's $g = 0.50$. This means that only a minority of the non-normative answers can be explained by miscalculations and, possibly, by a lack of attention to the problem. To ensure no carry-over effect, we also tested any potential differences in the numerical task according to the wording condition of the Bayesian word problem. The idea behind such a test relied on the fact that the standard wording yielded close-to-zero performance, so no beneficial carry-over effect would be possible in that condition. Any positive carry-over effect to the numerical task would originate from the benefits of successfully solving the problems with the improved wording. However, we found no statistically significant differences when comparing the Bayesian performance across the conditions: a version of the problem with the standard wording, the improved wording, the reworded question and the reworded information description (88.5%, 88.0%, 81.6%, 88.8%, respectively), $\chi^2(3) = 6.68$, $p = .083$, Cramer's $V = 0.09$. Thus, the absence of statistical differences across the differently worded problems undermines the possibility of a carry-over effect.

To conclude, we found strong support for the wording effect in the word problems featuring the natural frequency format. The wording effect was mainly driven by how the verbal description of the statistical information in the problem was worded. We have gathered additional evidence supporting the notion that the considerable disparity in Bayesian performance can be primarily attributed to differences in understanding the problem, rather than being solely driven by arithmetic difficulties.

7. Experiment 6

In Experiment 6, we aimed to replicate the wording effect of the Trisomy 21 problem using different numerical information that would result in a much higher conditional probability of around 31%. It is not clear whether our previous findings generalised to different levels of probability/frequency. In addition, we tested whether the wording effect would translate into normalised frequencies using percentages. On one hand, combining normalised frequencies is computationally much more challenging than combining natural frequencies (Ayal & Beyth-Marom, 2014). This is because the normalisation of numbers requires participants to multiply the base rate information with the diagnostic information (i.e., P(H)* P(D|H) and p(¬H)*p(¬D|¬H)) before generating the required ratio. On the other hand, prior research showed that the active ingredient of the facilitatory effect of the natural frequencies, relative to the normalised frequencies, might have been the textual partitive structure rather than the numbers themselves (Macchi, 2000). However, relatively small numbers of participants per group were used in this research, which might hinder any potential differences, especially when the performance was low. Therefore, we tested whether a clearer description of the structure of the word problem would lead to better performance, regardless of the statistical format. Based on these considerations, we hypothesised that the improved wording of the problems featuring natural frequencies would increase Bayesian reasoning relative to the standard wording. We also hypothesised that participants would perform better with the improved wording with the natural frequencies than with the normalised frequencies. Finally, we hypothesised that participants would perform better with the improved wording of the normalised frequencies than with the natural frequencies using the standard wording.

7.1. Method

7.1.1. Participants and design

We aimed to recruit at least 510 participants based on an a-priori power analysis. For each comparison, 169 participants were needed in each group to detect the small-to-medium effect of $w = 0.20$ while using a chi-squared test of independence to test the hypotheses and assuming $\alpha = .017$, $1 - \beta = .90$, a two-sided test and a small attrition rate (Cohen, 1988). We believed that the assumed effect size would be a conservative estimate given that we found $w = 0.29$ and $w = 36$ with the Trisomy 21 problem in Experiments 4 and 5. Therefore, 510 participants were recruited from the online panel, Prolific, and completed an online questionnaire. None of the participants was excluded based on preregistered criteria. The sample was balanced in terms of participants' sex. Only participants who were at least 18 years old, and UK nationals currently residing in the UK, with at least a 90% approval rate from previous studies were eligible to participate. The participants were paid £0.30 for completing a 3-min questionnaire.

Participants' ages ranged from 19 to 80 years ($M = 41.3$, $SD = 13.5$ years), and 50.6% were men, 48.6% were women and 0.8% were of other gender. The levels of education achieved by participants were relatively heterogeneous: 0.6% did not complete their high school education, 35.5% completed high school education, 44.7% completed an undergraduate degree, 16.5% completed a master's degree and 2.7% completed a PhD or other professional degree. The sample consisted of managers and working professionals (32.7%), unemployed people, students and homemakers (13.7%); workers in sales and offices (11.6%), retired (6.9%), workers in services (6.5%), in government (6.5%) or some other occupation category.

In a simple between-subjects design, participants solved a medical screening scenario presented either as natural frequencies using standard wording ($n = 174$), natural frequencies using improved wording ($n = 172$) or normalised frequencies using standard wording ($n = 164$). They were allocated to conditions randomly.

7.1.2. Materials and procedure

After providing informed consent, participants read and solved the Trisomy 21 problem describing a neck-fold prenatal test as an indicator of a child with Down syndrome (Galesic et al., 2009). Participants then estimated the number (or percentage) of women having a child with Down syndrome out of those with a positive screening test. The standard and improved wording conditions for the natural frequency format were identical to those used in Experiments 4 and 5, with two exceptions. First, the numerical information was changed for both conditions. We used a higher base rate (10 out of 100), the same false-positive rate (8 out of 10) and a higher false-negative rate (18 out of 90). This was done to test a higher level of conditional probability (8 out of 26) while using easy-to-compute numbers in all three conditions. Second, in the improved wording condition, we removed the repeated mention of the number of true-positives from the phrase introducing false-positives ("In addition to these 8 women …") to achieve equivalence of wording across the conditions. The normalised frequencies using the improved wording used the same wording and equivalent numerical information but were normalised over 100, i.e., using percentages. (Please see Supplementary Material for the exact wording of the problems.) The numbers were chosen to make the mental calculation relatively easy; for example, "80% percent out of 10%" and "20% out of 90%" are easy to compute mentally. None of the presented and calculated numbers used decimal places. Participants then estimated the number/percent of women having a child with Down syndrome out of those with a positive screening test. Afterwards, the participants answered questions concerning their age, gender, level of education and occupation.

As in the previous experiments, we used a strict criterion to code the normatively correct answers corresponding to the Bayesian calculations (i.e., $8/(8 + 18)$ and the absolute deviation from the normatively correct answer as a secondary measure. We used the same statistical analyses as in previous experiments to test the hypotheses.

7.2. Results and discussion

Participants provided more Bayesian answers when presented with

the improved wording of the question, the description of the statistical information and both, compared to the standard wording of the natural frequency version of the task, which was close to zero (see Table 2). The performance, however, varied substantially across the improved wording conditions. The statistically significant omnibus test of differences, $\chi^2(2) = 42.05$, $p < .001$, Cramer's $V = 0.29$, was followed up by pairwise comparisons to test our hypotheses (using the Bonferroni correction $\alpha = .05/3 = .017$). As predicted, the participants generated significantly more Bayesian answers in the natural frequency problem with the improved wording than with the standard wording, $\chi^2(1) = 24.51$, $p < .001$, Cramer's $V = 0.27$, yielding $BF_{10} = 95.6*10^3$ to 1 in favour of the association between the two different types of wording and the number of correct responses. The participants generated roughly the same number of Bayesian answers in the natural frequency problem with the standard wording compared to the normalised frequencies with the improved wording, $\chi^2(1) < 0.01$, $p = 1.000$, Cramer's $V < 0.01$, yielding $BF_{01} = 18.1$ to 1 in favour of no association between the different conditions and the number of correct responses. Finally, the participants generated significantly more Bayesian answers in the natural frequency problem with the improved wording than with the normalised frequencies with the improved wording, $\chi^2(1) = 22.54$, $p < .001$, Cramer's $V = 0.26$, yielding $BF_{10} = 35.3*10^2$ to 1 in favour of the association between the two different types of wording and the number of correct responses. Thus, we confirmed the hypotheses about the

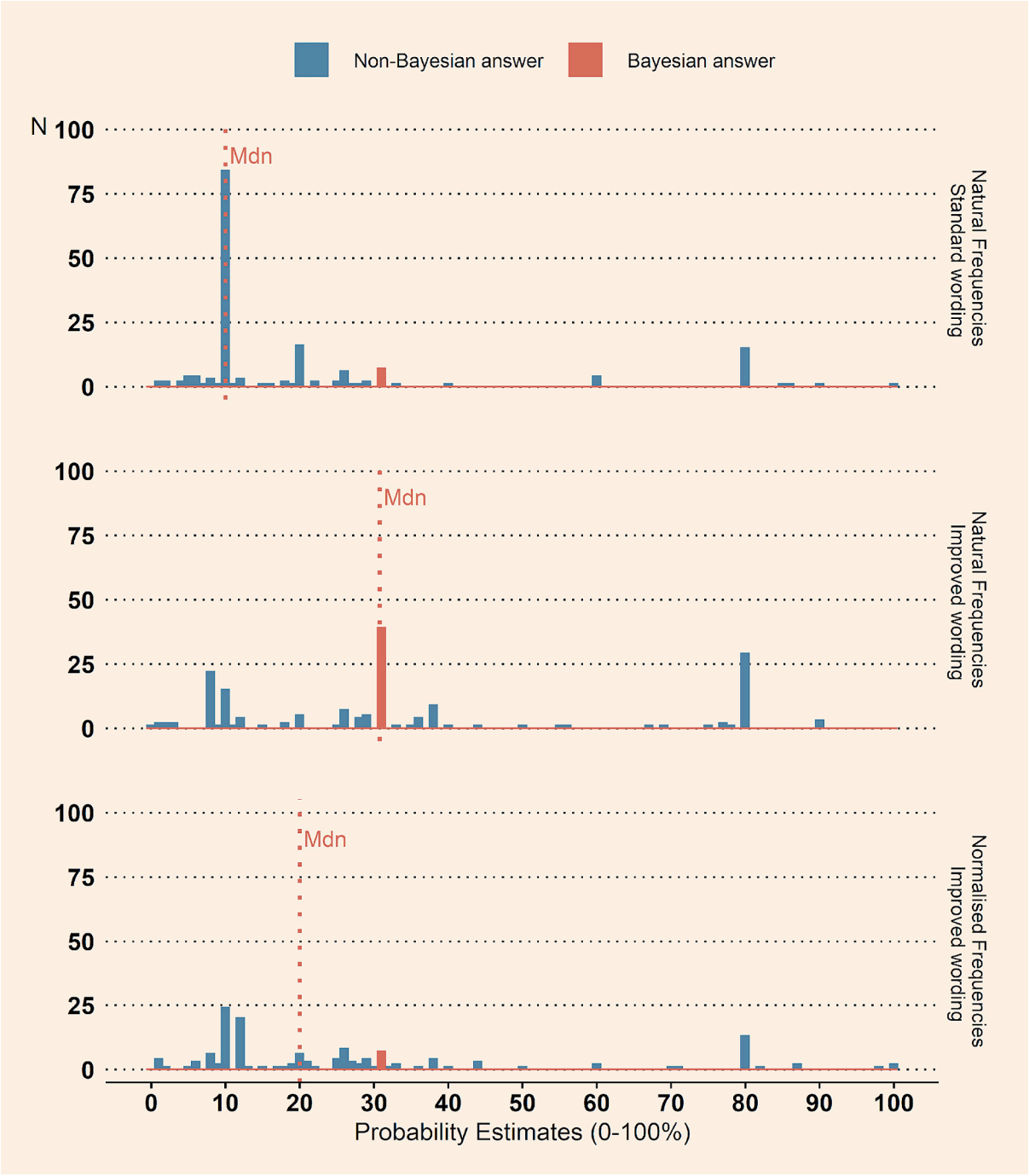

Fig. 6. Effect of statistical format and problem-wording on Bayesian reasoning.
The dotted vertical line represents the median probability estimate (*Mdn*) within each condition.

improved wording effect with the natural frequencies as compared to the standard wording with the natural frequencies and the normalised frequencies with the improved wording. However, we did not confirm the effect of the improved wording of the normalised frequencies.

Overall, participants deviated somewhat from the Bayesian estimate with some limited variability across the conditions (Table 2; Fig. 6). We had to remove the estimates from 22 participants from the normalised frequency condition as they were yielding probabilities higher than 100%. (We still coded these as incorrect responses in the main variable.) We found a statistically significant omnibus difference between the conditions, *K-W*(2) = 6.61, $p = .037$. The omnibus difference was followed up by pairwise comparisons to test our hypotheses (with an adjusted $\alpha = .017$). We found insufficient support for the difference between the standard and improved wording in the natural frequency format, $W = 16{,}752$, $p = .051$, yielding $BF_{01} = 1.1$ to 1 in favour of no difference according to the wording in the absolute deviation from the normatively correct answer. Similarly, we found insufficient support for the difference between the standard wording natural frequencies and the normalised frequencies with the improved wording, $W = 14{,}346$, $p = .012$, yielding $BF_{01} = 1.2$ to 1 in favour of no difference according to the wording in the absolute deviation from the normatively correct answer. Lastly, we found some support for the lack of difference between the natural frequencies and the normalised frequencies with the improved wording, $W = 11{,}613$, $p = .453$, yielding $BF_{01} = 3.9$ to 1 in favour of no difference between the two formats featuring the improved wording in the absolute deviation from the normatively correct answer. Thus, given the anecdotal evidence, we could not (dis-)confirm the hypothesised differences between the conditions except for some support of no difference between the two formats featuring the improved wording.

To conclude, we found strong support for the wording effect in the word problems featuring the natural frequency format even with higher-level probabilities. The higher base rate of having a child with Down syndrome is realistic for pregnant women between 48 and 50 years (Pighin et al., 2015). We did not explicitly specify the age group in the scenario to allow for a direct comparison of the scenarios across the experiments. However, including a reference group within a specific age range would enhance the ecological validity of the scenario. The improved wording did not yield the expected improvement with the normalised frequencies. The performance with the normalised frequencies was similar to the performance with the natural frequencies using the standard wording.

8. Experiment 7

In Experiment 7, we used the format of single-event probabilities and natural frequencies, for which the wording was varied as well while using low prior and high posterior probabilities. Whereas the prior and posterior probabilities in all scenarios tested before are clearly separated on the implied frequency scale—the ratio of posterior to prior probability is around two in Experiments 1–3, around ten in Experiments 4–5, and around three in Experiment 6—it might be hard to see such a separation on the 0–1 probability scale for some of them. For instance, in Experiment 4, the base rate was 15 out of 10,000, whereas the Bayesian answer was, ten times higher, 12 out of 811; however, on the probability scale these values appear close to each other: 0.2% vs 1.5%. Thus, in Experiment 7, we tested the wording and natural frequency effect with a low prior (1%) and high posterior probability (29%), with a ratio of 29. This should complement the evidence of Experiment 6 (prior of 10% and posterior of 31%) demonstrating that the standard wording, in contrast with the improved wording, yielded mostly base-rate only answers, a few Bayesian answers, and a lack of the natural frequency effect.

Moreover, we made two changes to improve absolute performance and avoid floor effects that might hinder the facilitating effect of natural frequencies. First, while using the wording of the Trisomy 21 problem, we replaced the medical terminology, possibly taxing the working memory of participants, with a generic description (i.e., genetic illness, genetic test). Second, we increased the financial reward participants received for participating in the research. Prior research showed that participants allocate cognitive effort to solving verbal problems based on cost and benefit considerations (Sirota, Juanchich, & Holford, 2023).

We hypothesised that natural frequencies in standard wording and improved wording would facilitate Bayesian reasoning relative to the single-event probability format. More importantly, we hypothesised that the improved wording would have an additional facilitative effect for problems featuring natural frequencies.

8.1. Method

8.1.1. Participants and design

We aimed to recruit 600 participants based on the same stopping rule and power analysis as reported in Experiment 1. Using the apriori exclusion criteria, none of the participants was excluded. The analytical sample size was $N = 601$. The participants were recruited from an online UK panel (Prolific) using the same eligibility criteria as in Experiment 1. The participants were paid £0.40 for completing a 3-min questionnaire.

Participants' ages ranged from 18 to 77 years ($M = 43.6$, $SD = 13.4$ years); 50.2% were women, 48.9% were men and 0.8% were of other gender identities. The levels of the participants' education were relatively heterogeneous: 1.0% did not complete their high school education, 33.8% completed high school education, 44.8% completed a college degree, 17.6% completed a master's degree and 2.8% completed a PhD or other professional degree. The sample consisted of managers and working professionals (36.4%), retired (11.5%), workers in sales and offices (11.3%), unemployed people, students and homemakers (9.6%), service workers (7.2%), government workers (6.2%) and some other, less common, occupations.

In a simple between-subjects design, participants were randomly allocated to one of the three conditions: single-event probability ($n = 208$), natural frequencies using standard wording ($n = 199$) and natural frequencies using improved wording ($n = 194$). Participants estimated the probability or frequency of a woman having a child with a genetic illness given the positive genetic test.

8.1.2. Materials and procedure

The procedure was identical to that in Experiment 1. The problem was adopted verbatim from the Trisomy 21 problem used in prior experiments but featured two changes (see Supplementary Materials). First, the prior probability (100 out of 10,000 women; 1%) and posterior probability (80 out of 278; 29%) were changed to be able to unequivocally differentiate strictly correct answers from those base rate-only responses. Second, the content was slightly modified: Trisomy 21 was replaced with "genetic illness" and the "neck-fold test" was replaced with a "genetic test". In addition, we asked participants to imagine a new sample (i.e., "Imagine a new representative sample…" instead of "Here is a new representative sample…") to avoid possible confusion about the new sample.

We used the strict coding of Bayesian answers and the calculation of the absolute deviation from the correct answer as before. We followed the pre-registered analytical strategy, which was identical to that used in prior experiments.

8.2. Results and discussion

Across the conditions, only a minority of participants calculated the Bayesian answers. We observed the same pattern as before across the conditions (see Table 2): a few Bayesian answers in the probability and natural frequency conditions using the standard wording and an increased number of Bayesian answers using the improved wording (see Table 2). The omnibus test confirmed the existence of statistically significant differences between the three conditions, $\chi^2(2) = 32.32$, $p < .001$, Cramer's $V = 0.23$.

To test our hypotheses, we conducted three pairwise comparisons (we used the Bonferroni correction and adjusted $\alpha = .05/3 = .017$). We found no significant difference between the single-event probability condition and the natural frequency condition with standard wording, $\chi^2(1) = 4.37$, $p = .037$, Cramer's $V = 0.10$, with $BF_{01} = 1.5$ to 1, providing only anecdotal evidence favouring slightly no association between the statistical format and the number of correct responses. The improved wording of the natural frequency format yielded significantly more Bayesian answers than the single-event probability one, $\chi^2(1) = 24.73$, $p < .001$, Cramer's $V = 0.25$, yielding $BF_{10} = 28.2*10^4$ to 1 in favour of the association between the two types of problem and the number of correct responses. Finally, the improved wording compared to the standard wording also facilitated Bayesian reasoning in the problems featuring natural frequencies, $\chi^2(1) = 9.79$, $p = .002$, Cramer's $V = 0.16$, yielding $BF_{10} = 18.5$ to 1 in favour of the association between the two different types of wording and the number of correct responses.

In terms of deviation from the Bayesian estimate, we found considerable variability across the conditions (Table 2; Fig. 7). After removing one implausible value exceeding 100%, we found a statistically significant omnibus difference between the conditions, $K\text{-}W(2) = 85.99$, $p < .001$. In pairwise comparisons to test our hypotheses (with an adjusted $\alpha = .017$), we noted less deviation with the improved wording compared to the standard wording in the natural frequency format, $W = 24{,}392$, $p < .001$, yielding $BF_{10} = 70.5*10^1$. Additionally, the improved wording

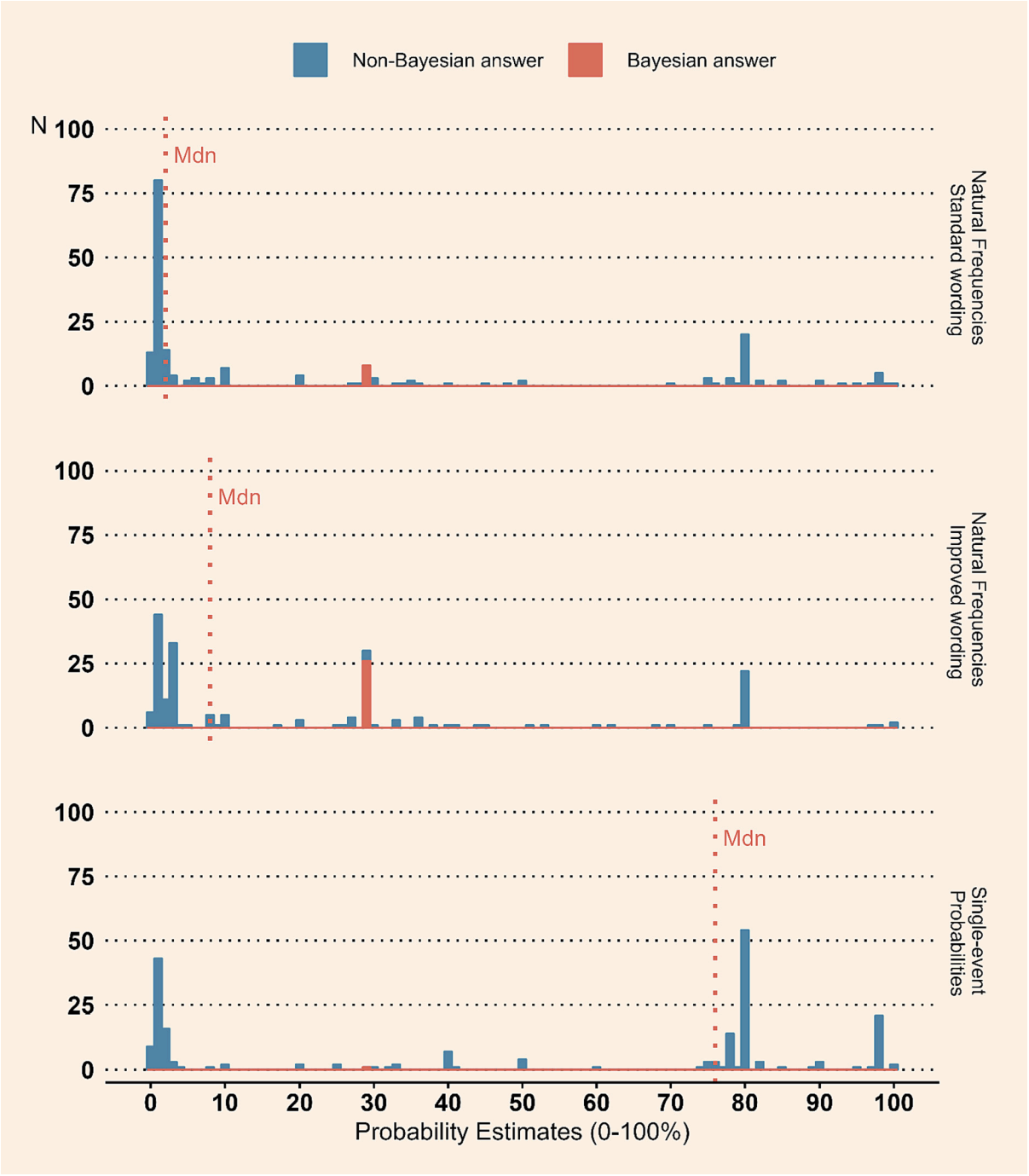

Fig. 7. Effect of statistical format and problem-wording on Bayesian reasoning in experiment 7.
The dotted vertical line represents the median probability estimate (*Mdn*) within each condition.

of the natural frequencies resulted in less deviation from the correct answer compared to the single-event probabilities, $W = 9838$, $p < .001$, yielding $BF_{10} = 17*10^8$. Finally, the standard wording of natural frequencies showed less deviation than the single-event probabilities, $W = 14{,}217$, $p < .001$, yielding $BF_{10} = 41.1*10^2$.

Thus, natural frequencies facilitated Bayesian reasoning compared to single-event probabilities, but this was only evident when we used the problem with improved wording. The improved (versus standard) wording of the problem also enhanced reasoning in problems featuring natural frequencies.

9. The overall wording effect

Across the seven experiments reported here, the improved wording (as opposed to standard wording) of the problem featuring natural frequencies has, on average, increased Bayesian reasoning by 14.6%. The meta-analytical odds ratio was $OR = 10.6$, 95% CI[5.6, 20.1], $z = 7.23$, $p < .001$, which is equivalent to Cohen's $d = 1.3$. Conventionally, this is considered to be a large effect size. It means that participants were more than ten times more likely to provide Bayesian answers with improved wording compared with the standard wording of the problems featuring the same statistical information in the natural frequencies format. To contextualise this evidence: the natural frequencies effect yielded, on average, a 20% increase in Bayesian reasoning with an odds ratio of 7.1 in the most comprehensive meta-analysis of the effect to date (McDowell & Jacobs, 2017). Therefore, the wording effect can be considered both strong and robust.

10. General discussion

In seven well-powered experiments, we found that the wording of the problem enhanced the Bayesian performance. The wording effect was reliably replicated with a different medical scenario (Experiments 3–7), and it was mainly driven by the rewording of the statistical information provided about the true and false-positives (Experiment 4). Additionally, we found that members of the public struggled to arrive at normatively correct responses when presented with single-event probabilities or normative frequency formats (Experiments 1, 3, 6 and 7), but also with natural frequencies using the standard wording. We did not find the facilitatory effect of the natural frequencies across these studies if the standard wording was used. Thus, the combination of the easier-to-calculate statistical format and the wording of the problem are the two important ingredients of the facilitatory effect of the natural frequencies—the finding that was mostly overlooked in prior research.

The critical observation of the wording effect is aligned with the prior literature on the importance of language comprehension in the mathematical problem-solving literature (Fuchs et al., 2015; Gros et al., 2020; LeBlanc & WeberRussell, 1996; Strohmaier et al., 2022) as well as the scarce evidence of the direct effect of wording on Bayesian reasoning with textbook problems (e.g., Johnson & Tubau, 2013). Indeed, the improved wording must have enabled problem-comprehension. Aligned with this finding, people did not have a problem with the calculations per se, as their performance skyrocketed when asked to compute equivalent arithmetic expressions in Experiment 5. The wording effect was mainly driven by the verbal description of the statistical information. The huge performance gap between the performance is further evidence that problem-comprehension is critical for Bayesian textbook problems. To avoid possible confusion among our readers, we do not claim that the wording used here is the most effective wording for Bayesian screening problems. More effective rewording of the problems might exist. However, for the sake of comparisons with the standard wording, and across the experiments, we kept the improved wording constant. Future research might identify a more optimal rewording of these problems that could serve as a standardisation example.

10.1. Theoretical and methodological implications of the wording effect

The wording effect has several theoretical and methodological implications. First, in past research, the wording of Bayesian problems differed between different statistical formats, and thus the wording represented a possible confounding variable in between-formats comparisons. (E.g., when comparing natural frequencies with single-event probability formats.) Second, the wording of the problems featuring natural frequencies across different studies often varied considerably, so the wording might moderate the within-formats variability of the performance with natural frequencies. Finally, the wording effect can prompt the specification of the current theories of Bayesian reasoning.

First, the statistical format of natural frequencies is often compared with the format of single-event probabilities (e.g.,Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). However, when examining the condition, we can observe that the statistical format is not the only thing manipulated; typically, the problem-wording also differs between formats. For instance, in the seminal paper demonstrating the natural frequency effect, the standard format of single-event probabilities uses conditional if-statements to communicate basic statistical information. (E.g., "If a woman has breast cancer, the probability is 80% that she will get a positive mammogram.") However, this is not the case with the natural frequency format (e.g., "8 of every 10 women with breast cancer will get a positive mammogram"). Thus, problem-wording might be an important confounding variable in this research. More informative comparisons would use the same, or as close as possible, wording (e.g., "80% probability that a woman with breast cancer will get a positive mammogram"). Or they would use relative frequencies (e.g., "80% of women with breast cancer will get a positive mammogram"). To be clear, given the overwhelming evidence of prior research and some evidence presented here, we do *not* argue that natural frequencies do not facilitate Bayesian reasoning at all. Rather, we argue that future research should try to estimate the contribution of the problem-wording and the statistical format separately when comparing the natural frequency format with that of single-event probabilities.

Second, the wording of the problem might moderate the performance with the natural frequency format. Consistently, prior meta-analysis indicated a substantial variability in the performance with the natural frequencies (McDowell & Jacobs, 2017). Of course, several variables, such as methodological differences and individual characteristics, can account for some of the variability (McDowell & Jacobs, 2017). Still, the wording of the problems might be one of the hidden moderators not systematically investigated. For example, consider the following snippet of the wording of the textbook problems featuring natural frequencies where adults and even sixth-graders achieved high scores:

> "…Of the 10 people who lie, 8 have a red nose. Of the remaining 90 people who don't lie, 9 also have a red nose." (Zhu & Gigerenzer, 2006, p. 297).

And compare this with the phrasing of the same type of information communicated in the textbook problems used in prior research (Galesic et al., 2009) and used here:

> "…If a person has insulin-dependent diabetes, it is not sure that he or she will have a positive result on the genetic test. More precisely, 48 of every 50 of such people will have a positive result on the genetic test. If a person does not have insulin-dependent diabetes, it is still possible that he or she will have a positive result on the genetic test. More precisely, 4,975 out of every 9,950 such people will have a positive result on the genetic test."

So, there are stark differences in the wording of the problem communicating similar statistical information in the research investigating the effect of natural frequencies. Future studies investigating the effects of statistical format should *standardise* the wording in each task to avoid introducing wording as a hidden moderator.

Finally, the effect of wording points out that the currently dominating theories of Bayesian problem-solving—the theories adopting the ecological rationality view (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995) or nested set theories (Barbey & Sloman, 2007; Girotto & Gonzalez, 2001)—are under-specified. Other factors contribute to the Bayesian performance of the problems featuring natural frequencies, which require a richer theoretical explanation. So, future versions of these theories could specify how they can account for and integrate other effects involved in Bayesian problem-solving such as the wording effects reported here, and the question format effects reported before (Girotto & Gonzalez, 2001). Alternatively, future theories could be developed building on the frameworks adopted from the literature on problem-solving. These could then account for a wider array of effects documented in the rich literature on Bayesian problem-solving. In fact, there have already been several calls (Johnson & Tubau, 2017; McNair, 2015; Sirota, Vallee-Tourangeau, et al., 2015) and attempts to develop such theoretical positions inspired by problem-solving literature (Tubau, 2021; Vallee-Tourangeau et al., 2015).

From a broader theoretical perspective, text comprehension and semantic clarity may be an additional explanatory mechanism behind the experience-description gap documented in probabilistic inferences and other domains (Hertwig & Erev, 2009; Lejarraga & Hertwig, 2021; Vance & Oaksford, 2021). In a nutshell, providing participants with a direct experience of relevant probabilities, as opposed to presenting them with summarised descriptions tends to reduce or mitigate various cognitive biases, including base rate neglect (Armstrong & Spaniol, 2017; Schulze & Hertwig, 2021). Previous explanations primarily focused on differences in the learning process—relying on direct experience of statistical information, as opposed to abstract, symbolic descriptions (e.g., Schulze & Hertwig, 2021)—when clarifying the gap. However, the semantic clarity of written descriptions may also contribute to creating the gap. Consider, for instance, the innovative experiment directly comparing experienced and descriptive statistical formats (Armstrong & Spaniol, 2017). Participants in the descriptive format condition received statistical information in an aggregated form (aggregated as natural frequencies), while those in the experienced format condition received sequentially presented information about individual patients. However, participants in the descriptive format condition must also read much more text than those in the experienced format—they need to comprehend the text to create an appropriate task representation. Thus, the participants differed not only in the way they acquired statistical information but also in the way they created a problem representation. This invites consideration of semantic factors to account for some of the observed differences. Such differences, in general, underscore the potential role of semantic clarity in influencing other cognitive biases. Notably, in the "heuristics and biases" research programme, participants were often tasked with solving verbal problems that may lack the clarity and context offered by real-life experiences (Tversky & Kahneman, 1974). This insight highlights the need for further investigation into how the comprehension of textual information can impact other cognitive and decision-making biases.

10.2. Practical implications

Regarding practical implications, our findings show that people from a general adult population find it difficult to draw normatively correct estimates from information about diagnostic test effectivity. Drawing on our findings and communication research literature in other domains (Fagerlin, Zikmund-Fisher, & Ubel, 2011), we recommend communicating the results of screening tests using natural frequencies but employing a simple and user-friendly language, which facilitates problem representation and, in turn, the correct calculation. We also recommend reaching out to other different representations of diagnostic test problems. These include distributive assessment using a different method of judgement elicitation (e.g., Pighin, Tentori, Savadori, & Girotto, 2018), and using visual representations of statistical information as these were shown to be very effective (Garcia-Retamero, Cokely, & Hoffrage, 2015; Garcia-Retamero & Hoffrage, 2013). However, given the recorded difficulties, the simplest method would be to avoid burdening patients and the public with any computations. Instead, provide them with posterior probability estimates directly with an appropriate explanation of how these estimates were reached (Navarrete, Correia, Sirota, Juanchich, & Huepe, D., 2015).

10.3. Effects of same sample type, causal explanation and statistical format

Besides the wording effect, three other findings of our research should be discussed in depth. The effect of the two additional interventions—the same sample type and causal explanation—did not create a substantial benefit compared with the control condition. Only one prior study demonstrated the positive effect of the same sample type (Johnson & Tubau, 2017). It, therefore, remains important to seek further replication and possible moderators of the effect. For the role of causal explanation, several prior studies yielded mixed findings, including positive and null effects of such a manipulation (Hayes, Ngo, Hawkins, & Newell, 2018; Krynski & Tenenbaum, 2007; McNair & Feeney, 2014, 2015). Our findings provide further evidence of the absence of the effect in relatively complex problems. Still, possible moderators of the effect should also be considered in future studies. It is possible that the screening tasks were so difficult in the standard wording for our participants that this prevented the manifestation of the facilitatory effects of those interventions. Indeed, the prior studies typically recorded a higher absolute performance in the control conditions and used university student samples.

Furthermore, our findings also demonstrated the close-to-zero performance with the natural frequencies and a lack of natural frequency effect, which seemingly contradicts the current literature. For instance, the recent meta-analysis reported a solid performance with natural frequencies and a robust effect of the natural frequency format compared to non-normalised formats such as single-event probabilities (McDowell & Jacobs, 2017). We believe the contradiction is illusory; in fact, our results might fit well with the currently available corpus of evidence and extend our understanding of it. First, the prior studies studying samples from a general adult population recorded a much lower normative performance with natural frequencies than those studying student or expert samples across different tasks (McDowell et al., 2018; McDowell & Jacobs, 2017). Some of them even reported a lack of the facilitatory effect of natural frequencies (Pighin et al., 2016). In our studies, we used online panel samples drawn from a general adult population, which could account for the overall lower performance in our studies. Second, the prior studies investigating medical textbook problems recorded much lower Bayesian performance than non-medical textbook problems (Siegrist & Keller, 2011). In our studies, we used medical textbook problems; hence, we expected the performance to be lower. The additive effect of these two factors, along with the difficult wording, might explain the flooring effect of the natural frequencies in our experiments resulting in the absence of their facilitatory effect while not putting into question the robust evidence of the general facilitative effect of natural frequencies (McDowell & Jacobs, 2017).

Different, non-exclusive mechanisms could be responsible for the low performance with natural frequencies reported here. For instance, the difference in general cognitive resources—between the student and non-student samples—as well as in cognitive requirements of the task—between the medical and non-medical problems—can account for this difference (De Neys, 2007; Lesage et al., 2013; Sirota, Juanchich, & Hagmayer, 2014). Prior research found that both single-event probabilities and natural frequencies recruit similar cognitive resources that are likely required to build proper problem representation and to perform correct computations (Sirota, Juanchich, & Hagmayer, 2014). In addition, student samples usually exhibit a higher numerical ability than non-student samples, which has been linked with superior

performance with natural frequencies (Sirota & Juanchich, 2011). Cognitive resource requirements can also explain the facilitative effect of the simplified wording. If reading comprehension recruits substantial executive resources, then making a word problem more comprehensible should free up some cognitive resources required for text comprehension. In turn, they can be used to build adequate problem representation and ease the required calculations (Kintsch, 1988; Kintsch & Greeno, 1985).

10.4. Limitations and future research

Three limitations of our research deserve more attention and should be addressed in future research. First, our samples from a general adult population provided a better picture of people's ability of probabilistic reasoning because they diverged from the majority of articles on Bayesian reasoning using only student samples. However, the samples we used were not probabilistically representative of a general adult population. Future research could therefore use probabilistic samples of general adult populations to estimate precisely how the public understands probabilities involved in interpreting the outcome of medical screening tests. Second, we replicated the wording effect on different medical textbook problems, but the effect might vary across contexts. Such findings, along with the evidence from mathematical problem-solving literature, point towards the generalisability of the effect. Nevertheless, future research should test the wording effect across other medical and non-medical textbook problems. Finally, the wording effect in our studies focused on changing the wording concerning statistical information and question format; however, a systematic exploration of other changes to the wording should be undertaken. This would be an essential avenue for any future research to identify the most effective communication strategies.

11. Conclusion

To conclude, the wording of Bayesian textbook problems matters. While the standard wording of the screening problems with natural frequencies in a general adult population did not improve Bayesian reasoning, relative to the single-event probabilities, the improved wording did. This research has identified a theoretically salient moderator of Bayesian reasoning with natural frequencies. The wording effect extends the current theoretical explanations of Bayesian reasoning and aligns with the recent focus on problem-solving literature. Ultimately, even intuitive statisticians must be good readers when solving Bayesian textbook problems.

CRediT authorship contribution statement

Miroslav Sirota: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Visualization, Writing – original draft, Writing – review & editing. **Gorka Navarrete:** Data curation, Resources, Writing – review & editing. **Marie Juanchich:** Investigation, Resources, Writing – review & editing.

Declaration of competing interest

We declare no conflict of interest.

Gorka Navarrete was awarded a grant from the National Agency for Research and Development (ANID/ FONDECYT Regular 1211373) which did not affect the integrity of the presented research.

Data availability

The materials, data sets, codebook, R code and preregistrations are publicly available at Open Science Framework https://osf.io/kp3g7/

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2024.105722.

References

Armstrong, B., & Spaniol, J. (2017). Experienced probabilities increase understanding of diagnostic test results in younger and older adults. *Medical Decision Making, 37*(6), 670–679.

Ayal, S., & Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgment and Decision making, 9*(3), 226–242. https://doi.org/10.1017/S1930297500005775

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes [article]. *Behavioral and Brain Sciences, 30*(3), 241-+.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211–233.

Bramwell, R., West, H., & Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: Experimental study. *BMJ, 333*(7562), 284.

Brase, G. L., & Hill, W. T. (2017). Adding up to good Bayesian reasoning: Problem format manipulations and individual skill differences [article]. *Journal of Experimental Psychology. General, 146*(4), 577–591. https://doi.org/10.1037/xge0000280

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision making, 4*(1), 34–40.

Cohen, A. L., Sidlowski, S., & Staub, A. (2017). Beliefs and Bayesian reasoning [article]. *Psychonomic Bulletin & Review, 24*(3), 972–978. https://doi.org/10.3758/s13423-016-1161-z

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

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

Cummins, D. D. (1991). Children's interpretations of arithmetic word problems. *Cognition and Instruction, 8*(3), 261–289.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*(4), 405–438. https://doi.org/10.1016/0010-0285(88)90011-4

Davis-Dorsey, J., Ross, S. M., & Morrison, G. R. (1991). The role of rewording and context personalization in the solving of mathematical word problems. *Journal of Educational Psychology, 83*(1), 61.

De Neys, W. (2007). Nested sets and base-rate neglect: Two types of reasoning? *Behavioral and Brain Sciences, 30*(3), 260–261.

Edwards, W., Lindman, H., & Phillips, L. D. (1965). *Emerging technologies for making decisions*.

Fagerlin, A., Zikmund-Fisher, B. J., & Ubel, P. A. (2011). Helping patients decide: Ten steps to better risk communication. *JNCI: Journal of the National Cancer Institute, 103*(19), 1436–1443. https://doi.org/10.1093/jnci/djr318

Fuchs, L. S., Fuchs, D., Compton, D. L., Hamlett, C. L., & Wang, A. Y. (2015). Is word-problem solving a form of text comprehension? *Scientific Studies of Reading, 19*(3), 204–223. https://doi.org/10.1080/10888438.2015.1005745

Galesic, M., Gigerenzer, G., & Straubinger, N. (2009). Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Medical Decision Making, 29*(3), 368–371. https://doi.org/10.1177/0272989x08329463

Garcia-Retamero, R., Cokely, E. T., & Hoffrage, U. (2015). Visual aids improve diagnostic inferences and metacognitive judgment calibration. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00932. Article 932.

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

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

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form [article]. *Cognition, 78*(3), 247–276. https://doi.org/10.1016/s0010-0277(00)00133-5

Glenberg, A., Willford, J., Gibson, B., Goldberg, A., & Zhu, X. (2012). Improving Reading to improve math. *Scientific Studies of Reading, 16*(4), 316–340. https://doi.org/10.1080/10888438.2011.564245

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*(9), 767–773.

Gros, H., Thibaut, J. P., & Sander, E. (2020). Semantic congruence in arithmetic: A new conceptual model for word problem solving. *Educational Psychologist, 55*(2), 69–87. https://doi.org/10.1080/00461520.2019.1691004

Hadianto, D., Damaianti, V. S., Mulyati, Y., Sastromiharjo, A., & Iop. (2020). Jul 14-15. In *Does reading comprehension competence determine level of solving mathematical word problems competence?Journal of physics conference series [international conference on mathematics and science education (icmsce) 2020]. International Conference on Mathematics and Science Education (ICMScE)*. Electr Network.

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

Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology, 87*(1), 18.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences, 13*(12), 517–523. https://doi.org/10.1016/j.tics.2009.09.004
Hoffrage, U., & Gigerenzer, G. (2004). How to improve the diagnostic inferences of medical experts. In *Experts in science and society* (pp. 249–268). Springer.
Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science, 290*(5500), 2261–2262.
Johnson, E. D., & Tubau, E. (2013). Words, numbers, & numeracy: Diminishing individual differences in Bayesian reasoning. *Learning and Individual Differences, 28*, 34–40. https://doi.org/10.1016/j.lindif.2013.09.004
Johnson, E. D., & Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00938. Article 938.
Johnson, E. D., & Tubau, E. (2017). Structural mapping in statistical word problems: A relational reasoning approach to Bayesian inference. *Psychonomic Bulletin & Review, 24*(3), 964–971. https://doi.org/10.3758/s13423-016-1159-6
Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review, 106*(1), 62–88. https://doi.org/10.1037/0033-295X.106.1.62
Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237.
Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163.
Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92*(1), 109.
Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology. General, 136*(3), 430–450. https://doi.org/10.1037/0096-3445.136.3.430
LeBlanc, M. D., & WeberRussell, S. (1996). Text integration and mathematical connections: A computer model of arithmetic word problem solving. *Cognitive Science, 20*(3), 357–407. https://doi.org/10.1207/s15516709cog2003_2
Leiss, D., Plath, J., & Schwippert, K. (2019). Language and mathematics - key factors influencing the comprehension process in reality-based tasks. *Mathematical Thinking and Learning, 21*(2), 131–153. https://doi.org/10.1080/10986065.2019.1570835
Lejarraga, T., & Hertwig, R. (2021). How experimental methods shaped views on human competence and rationality. *Psychological Bulletin, 147*(6), 535.
Lesage, E., Navarrete, G., & De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: The role of general cognitive resources. *Thinking & Reasoning, 19*(1), 27–53. https://doi.org/10.1080/13546783.2012.713177
Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior and Human Decision Processes, 82*(2), 217–236. https://doi.org/10.1006/obhd.2000.2895
Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul., 8*(1), 3–30. https://doi.org/10.1145/272991.272995
McDowell, M., Galesic, M., & Gigerenzer, G. (2018). Natural frequencies do Foster public understanding of medical tests: Comment on Pighin, Gonzalez, Savadori, and Girotto (2016). *Medical Decision Making, 38*(3), 390–399. https://doi.org/10.1177/0272989x18754508
McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin, 143*(12), 1273–1312. https://doi.org/10.1037/bul0000126
McNair, S., & Feeney, A. (2014). When does information about causal structure improve statistical reasoning? *Quarterly Journal of Experimental Psychology, 67*(4), 625–645. https://doi.org/10.1080/17470218.2013.821709
McNair, S., & Feeney, A. (2015). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychonomic Bulletin & Review, 22*(1), 258–264. https://doi.org/10.3758/s13423-014-0645-y
McNair, S. J. (2015). Beyond the status-quo: Research on Bayesian reasoning must develop in both theory and method. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00097. Article 97.
Morey, R. D., & Rouder, J. N. (2015). BayesFactor: An R package for Bayesian data analysis. In (version 0.9.10-2).
Navarrete, G., Correia, R., & Froimovitch, D. (2014). Communicating risk in prenatal screening: The consequences of Bayesian misapprehension. *Frontiers in Psychology, 5*, 1272. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4222132/pdf/fpsyg-05-01272.pdf.
Navarrete, G., Correia, R., Sirota, M., Juanchich, M., & Huepe, D.. (2015). Doctor, what does my positive test mean? From Bayesian textbook tasks to personalized risk communication [perspective]. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01327
Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon mechanical Turk. *Behavior Research Methods, 46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y
Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*(1), 29.
Pighin, S., Girotto, V., & Tentori, K. (2017). Children's quantitative Bayesian inferences from natural frequencies and number of chances. *Cognition, 168*, 164–175. https://doi.org/10.1016/j.cognition.2017.06.028
Pighin, S., Gonzalez, M., Savadori, L., & Girotto, V. (2016). Natural frequencies do not Foster public understanding of medical test results. *Medical Decision Making, 36*(6), 686–691. https://doi.org/10.1177/0272989x16640785
Pighin, S., Savadori, L., Barilli, E., Galbiati, S., Smid, M., Ferrari, M., & Cremonesi, L. (2015). Communicating down syndrome risk according to maternal age: "1-in-X" effect on perceived risk. *Prenatal Diagnosis, 35*(8), 777–782. https://doi.org/10.1002/pd.4606
Pighin, S., Tentori, K., & Girotto, V. (2017). Another chance for good reasoning [article]. *Psychonomic Bulletin & Review, 24*(6), 1995–2002. https://doi.org/10.3758/s13423-017-1252-5
Pighin, S., Tentori, K., Savadori, L., & Girotto, V. (2018). Fostering the understanding of positive test results. *Annals of Behavioral Medicine, 52*(11), 909–919. https://doi.org/10.1093/abm/kax065
Schulze, C., & Hertwig, R. (2021). A description–experience gap in statistical intuitions: Of smart babies, risk-savvy chimps, intuitive statisticians, and stupid grown-ups. *Cognition, 210*, Article 104580. https://doi.org/10.1016/j.cognition.2020.104580
Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130*(3), 380–400.
Siegrist, M., & Keller, C. (2011). Natural frequencies and Bayesian reasoning: The impact of formal education and problem context. *Journal of Risk Research, 14*(9), 1039–1055.
Sirota, M., & Juanchich, M. (2011). Role of numeracy and cognitive reflection in Bayesian reasoning with natural frequencies. *Studia Psychologica, 53*(2), 151–161.
Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning [article]. *Psychonomic Bulletin & Review, 21*(1), 198–204. https://doi.org/10.3758/s13423-013-0464-6
Sirota, M., Juanchich, M., & Holford, D. L. (2023). Rationally irrational: When people do not correct their reasoning errors even if they could. *Journal of Experimental Psychology. General, 152*(7), 2052–2073. https://doi.org/10.1037/xge0001375
Sirota, M., Kostovicova, L., & Juanchich, M. (2014). The effect of iconicity of visual displays on statistical reasoning: Evidence in favor of the null hypothesis [article]. *Psychonomic Bulletin & Review, 21*(4), 961–968. https://doi.org/10.3758/s13423-013-0555-4
Sirota, M., Kostovicova, L., & Vallee-Tourangeau, F. (2015). How to train your Bayesian: A problem-representation transfer rather than a format-representation shift explains training effects. *Quarterly Journal of Experimental Psychology, 68*(1), 1–9. https://doi.org/10.1080/17470218.2014.972420
Sirota, M., Thorpe, A., & Juanchich, M. (2022). Explaining and reducing the public's expectations of antibiotics: A utility-based signal detection theory approach. *Journal of Applied Research in Memory and Cognition, 11*(4), 587–597. https://doi.org/10.1037/mac0000027
Sirota, M., Vallee-Tourangeau, G., Vallee-Tourangeau, F., & Juanchich, M. (2015). On Bayesian problem-solving: Helping Bayesians solve simple Bayesian word problems. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01141. Article 1141.
Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes, 91*(2), 296–309. https://doi.org/10.1016/S0749-5978(03)00021-9
Staub, F. C., & Reusser, K. (1995). The role of presentational structures in understanding and solving mathematical word problems. In Discourse comprehension: Essays in honor of Walter Kintsch. (pp. 285–305). Lawrence Erlbaum Associates, Inc.
Strohmaier, A. R., Reinhold, F., Hofer, S., Berkowitz, M., Vogel-Heuser, B., & Reiss, K. (2022). Different complex word problems require different combinations of cognitive skills. *Educational Studies in Mathematics, 109*(1), 89–114. https://doi.org/10.1007/s10649-021-10079-4
Tubau, E. (2021). Why can it be so hard to solve Bayesian problems? Moving from number comprehension to relational reasoning demands. *Thinking & Reasoning, 1-20*. https://doi.org/10.1080/13546783.2021.2015439
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293.
Vallee-Tourangeau, G., Abadie, M., & Vallee-Tourangeau, F. (2015). Interactivity fosters Bayesian reasoning without instruction. *Journal of Experimental Psychology. General, 144*(3), 581–603. https://doi.org/10.1037/a0039161
Vallée-Tourangeau, G., Sirota, M., Juanchich, M., & Vallée-Tourangeau, F. (2015). Beyond getting the numbers right: What does it mean to be a "successful" Bayesian reasoner? *Frontiers in Psychology, 6*.
Vance, J., & Oaksford, M. (2021). Explaining the implicit negations effect in conditional inference: Experience, probabilities, and contrast sets. *Journal of Experimental Psychology: General, 150*(2), 354–384. https://doi.org/10.1037/xge0000954
Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J. E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology, 28*(4), 409–426. https://doi.org/10.1080/01443410701708228
Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets? *Experimental Psychology, 50*, 97–106. https://doi.org/10.1026/1618-3169.50.2.97
Zhou, X. L., Li, M. Y., Li, L. N. A., Zhang, Y. Y., Cui, J. X., Liu, J., & Chen, C. S. (2018). The semantic system is involved in mathematical problem solving. *NeuroImage, 166*, 360–370. https://doi.org/10.1016/j.neuroimage.2017.11.017
Zhu, L. Q., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation [article]. *Cognition, 98*(3), 287–308. https://doi.org/10.1016/j.cognition.2004.12.003