BRIEF REPORT

Another chance for good reasoning

Stefania Pighin¹ · Katya Tentori¹ · Vittorio Girotto²

Published online: 6 March 2017
© Psychonomic Society, Inc. 2017

Abstract Disagreement on the “probability status” of chances casts doubt on Girotto and Gonzalez’s (2001) conclusion that the human mind can make sound Bayesian inferences involving single-event probabilities. The main objection raised has been that chances are de facto natural frequencies disguised as probabilities. In the present study, we empirically demonstrated that numbers of chances are perceived as being distinct from natural frequencies and that they have a facilitatory effect on Bayesian inference tasks that is completely independent from their (minor) frequentist readings. Overall, therefore, our results strongly disconfirm the hypothesis that natural frequencies are a privileged cognitive representational format for Bayesian inferences and suggest that a significant portion of laypeople adequately handle genuine single-event probability problems once these are rendered computationally more accessible by using numbers of chances.

Keywords Chances · Single-event probability · Natural frequencies · Probabilistic reasoning · Bayesian reasoning

Prominent accounts of Bayesian reasoning generally agree that the accuracy of people’s probabilistic inferences depends on how the inferential problem is represented. However, these

accounts differ on what makes a representation able to improve performance.

According to the *frequentist* view (e.g., Gigerenzer & Hoffrage, 1995, 2007; Gigerenzer, Hoffrage, & Ebert, 1998; Hoffrage & Gigerenzer, 1998; Hoffrage, Krauss, Martignon, & Gigerenzer, 2015), Bayesian inferences are facilitated by a problem representation in terms of natural frequencies. Natural frequencies, it is argued, constitute a cognitively privileged format because they are the outcomes of natural sampling, the process of counting and appropriately classifying occurrences of events as they are encountered. On the other hand, the human mind “would not be tuned to probabilities or percentages as input format” (Gigerenzer & Hoffrage, 1995, p. 686), because these do not correspond to the typical way in which humans have handled statistical information during their evolution (see also Gigerenzer, 1996). The main support for such claims is provided by the experimental evidence that probability judgments are typically more accurate when problems are framed in terms of natural frequencies rather than percentages (see Gigerenzer & Hoffrage, 1995; Hoffrage & Gigerenzer, 2004).

According to the contrasting *nested-sets* view (e.g., Barbey & Sloman, 2007; Evans, Handley, Perham, Over, & Thompson, 2000; Fox & Levav, 2004; Girotto & Gonzalez, 2001; Johnson-Laird, Legrenzi, Girotto, Sonino Legrenzi, & Caverni, 1999; Sloman, Over, Slovak, & Stibel, 2003), the real source of facilitation is not the frequency format but the partition of the data into exhaustive subsets that explicitly provides responders with the relevant conjunctive events (referred to as *partitive* or *partitioned structure*; Girotto & Gonzalez, 2001; Macchi, 1995), together with a question form that prompts individuals to compute the two terms of the Bayesian ratio (e.g., the *two-step* question typically used in frequentist scenarios or, even better, the *distributive* question

✉ Stefania Pighin
stefania.pighin@unitn.it

¹ Center for Mind/Brain Sciences, University of Trento, Trento, Italy

² Department of Architecture and Arts, University IUAV of Venice, Venice, Italy

proposed by Girotto & Gonzalez, 2001; for more on these questions see the introduction of Exp. 2).

In clarification of their theoretical position, the defenders of the frequentist view (e.g., Gigerenzer & Hoffrage, 1995, 2007) specified that they did not mean to argue that the frequency information determines facilitation per se, but only that natural frequencies—unlike percentages or normalized frequencies—automatically incorporate base rate information and make the set relations explicit, with a consequent computational simplification.

Given that the two sides came to agree that the facilitation is not determined by natural frequencies per se, but rather by the structure and questions typically associated with them, it became crucial to see whether such structure and questions could also be implemented with single-event probabilities. Girotto and Gonzalez (2001) argued that this can be done by means of chances (which already appeared in Johnson-Laird et al., 1999). As well as natural frequencies, chances can state probabilities as positive integer numbers, can be arranged in subset relations, and can preserve the sample size of the reference class. Unlike natural frequencies, however, chances can express single-event probabilities. In three studies, Girotto and Gonzalez (2001) showed that, when the structure of the problem and the question were framed so as to prompt the use of extensional reasoning, individuals were able to solve Bayesian inference problems independently from the information type (chances vs. natural frequencies).

The advocates of the frequentist hypothesis rejected Girotto and Gonzalez's (2001) conclusion because they disagreed with the categorization of chances as single-event probabilities. According to this objection, since chances “are not single numbers in the interval $[0, 1]$ ” (i.e., they can be unnormalized), they are nothing but “natural frequencies disguised as probabilities” (Hoffrage, Gigerenzer, Krauss, & Martignon, 2002, p. 350; but see also Brase, 2002, and Gigerenzer & Hoffrage, 2007).

In their rejoinder to Hoffrage et al. (2002), Girotto and Gonzalez (2002) provided a theoretical argument in favor of the idea that chances are a standard single-event probability format that is not conflated with natural frequencies. In particular, they pointed out that chances do not necessarily rely on previous observations and are a common way to communicate probabilities that, historically, preceded the use of numbers in the $[0, 1]$ interval. That is, according to Girotto and Gonzalez (2002, p. 356), the expression “4 chances out of 52,” as referred to drawing a queen from a regular deck of cards, is not an expression of natural frequencies, because it does not mean that, in the past, a queen has been picked 4 times in a series of 52 draws, but only that there are “4 possibilities of drawing a queen out of a total set of 52 possibilities.” Accordingly, Girotto and Gonzalez (2002) argued that people still believe that the chances of drawing a queen from a regular deck are “4 out of 52” even if the frequencies they may have observed in

previous draws were different from the 4/52 ratio. This suggests that—unlike natural frequencies—chances are not necessarily the result of a sequential process of observing and counting past events, and also that—like proper single-event probabilities—chances can express theoretical possibilities that incorporate abstract knowledge (such as the information conveyed by the prior knowledge that the deck is a regular one).

In 2008, Brase moved the debate on the status of chances to an empirical level. In a series of three experiments, he found that, although the majority of respondents (63% across studies) interpreted chances as single-event probabilities, those who reported a frequentist interpretation tended to give more correct answers than did those who reported a single-event interpretation. Brase concluded that chances constitute “ambiguous statistical information” (p. 284) whose facilitation in statistical reasoning depends on their interpretation as frequencies.

Brase's (2008) conclusion has had a large impact on the literature (see Barton, Mousavi, & Stevens, 2007; Brase & Hill, 2015; Hill & Brase, 2012; Johnson & Tubau, 2015; Kynn, 2008; Moro, Bodanza, & Freidin, 2011). But is it empirically supported? Although the attempt to control experimentally how people interpret different representation formats is appreciable in many respects, in our opinion Brase's (2008) study cannot be considered conclusive because it suffers from serious methodological limitations. The first issue concerns the fact that, in Brase's (2008) experiments, only a minority of the participants (three, two, and six participants in Exps. 1, 2, and 3, respectively) correctly solved the chances version of the problem. As a consequence, the crucial comparison between the accuracy rates for participants who gave the single-event probability versus frequency interpretation of chances concerned two tiny subgroups (e.g., 1/18 vs. 2/9, 1/21 vs. 1/8, and 2/23 vs. 4/13 participants in Exps. 1, 2, and 3, respectively). As the author himself acknowledged (p. 286), such sample sizes preclude strong conclusions. The second concern (which might, in a sense, be the cause of the former) is that Brase's (2008) chance and frequency versions differed not only for the information conveyed but also for the question that was put to participants. In fact, whereas the frequency question closely resembled those more effectively employed in the literature, the chance question featured some peculiarities (both in the tenses and in the lack of reference to the overall set of 100 chances). This might explain why participants' success rate in Brase (2008) was much lower than in Girotto and Gonzalez (2001). The third methodological problem relates to the interpretations offered to participants in some of Brase's (2008) experiments. In particular, the expression “a large number of applications,” appearing in both the frequency and single-event probability interpretations of Experiment

2, could have led to confusion among the participants. Indeed, not only did one fourth of participants in the single-event probability condition choose the frequency interpretation, but even more participants (one third) in the natural-frequency condition chose the single-event probability interpretation. Our final (related) issue regards Brase's (2008) conclusion that his results proved that chances (but not frequencies) are ambiguous statistical information. Actually, a small-scale meta-analysis comparing the rates of chance and natural-frequency interpretations reported in his experiments did not provide evidence in favor of the hypothesis that chances are more ambiguous than natural frequencies (fixed-effects model, overall $OR = -0.26$, $CI = -0.94$ to 0.42 , $p = .449$).¹ All of the above remarks converge to suggest that Brase's (2008) study failed to prove that chances are ambiguous statistical information, and that it also failed to show any clear relation between participants' interpretations of chances and their reasoning performance.

Indirect evidence that a frequency interpretation of chances could facilitate Bayesian reasoning was provided by the empirical finding (Brase, 2009) that performances in problems featuring chances benefit more from iconic representations that mimic natural frequencies (i.e., pictograms) than from other graphical representations (i.e., Euler diagrams). However, this result was challenged by Sirota, Kostovičová, and Juanchich (2014), who claimed that it was merely an artifact generated by participants' expectations concerning the meaning of the different graphical representations and also by the specific wordings used to introduce these representations. On controlling for such aspects, in fact, they were unable to replicate the facilitatory effect of iconic representations reported in Brase (2009).

In a similar vein, Sirota, Kostovičová, and Vallée-Tourangeau (2015a) provided experimental evidence that a natural-frequency representation is not necessary for training to improve performance in Bayesian reasoning problems expressed in chances. According to their results, in fact, what is crucial for performance improvement is that participants learn to appreciate an adequate nested-sets structure.

Finally, Sirota, Kostovičová, and Vallée-Tourangeau (2015b) explored the possible mechanisms that could explain the gap between chances and natural frequencies reported by Brase (2008). Using Brase's (2008) versions of the Bayesian inference task, they replicated the greater facilitatory effect of natural frequencies than of chances. However, on using a different follow-up question, they did not find any support for the hypothesis that the mental representations of chances as

frequencies (rather than as probabilities) accounted for the format gap observed between chances and frequencies.

In the present study, we extend the above-mentioned empirical investigations on the controversial nature of chances. In what follows, we will describe two experiments aimed at both understanding how laypeople interpret chances and controlling whether their reasoning accuracy depends on how chances are interpreted. In Experiment 1, laypeople first had to interpret (as conveying single-event probability vs. frequency information) a statement expressed in one of various numerical formats, and then they had to answer a reasoning question aimed at assessing their understanding of that statement. In Experiment 2, laypeople first had to solve a Bayesian inference problem formulated in terms of numbers of chances, and then they had to indicate whether the information conveyed in the problem referred to either a single-event probability or a frequency.

Experiment 1

To investigate whether laypeople give different interpretations to chance and frequency statements conveying the same information and whether such interpretations affect their reasoning performance, we used two chance and two frequency statements concerning a possible infection and a test that was useful to detect it (see Table 1). The two chance statements were expressed as numbers of chances (*unnormalized chance* statement) and corresponding percentages (*normalized chance* statement), while the two frequency statements reported the numbers of observations of a previous sampling process (*natural frequency* statement²) and corresponding percentages (*normalized frequency* statement). We employed both unnormalized and normalized chance statements, as well as both natural and normalized frequency statements, in order to disentangle the effect of the (chance vs. frequency) format from that of the normalization of the numerical information.

Method

Participants The minimum sample size needed for Experiment 1 (as well as for Exp. 2) was computed by G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009). The power analysis showed that in order to detect at least a medium effect size ($\rho = 0.3$; Cohen, 1969, p. 76), assuming $\alpha = .05$ and $1 - \beta = .95$, the minimum sample required was 35 participants per group. Since we needed four groups of participants, we recruited 180 U.S. residents ($M_{age} = 36$ years, $SD = 11.7$; 98 men, 82 women) using the Amazon Mechanical Turk (AMT) platform. Most of the participants had a university

¹ We thank Miroslav Sirota for suggesting this analysis, which was performed using the "Metafor" package in the R statistical environment.

² Note that natural frequencies by definition are unnormalized.

Table 1 Statements and reasoning questions (with corresponding correct responses, within square brackets) used in Experiment 1

Numerical format	
Unnormalized chance	There are 5 chances out of 100 that a person has the infection. An infected person has 4 chances out of 5 of testing positive. What are the chances that a person has the infection and tests positive? <u>[4]</u> out of 100
Normalized chance	There is a 5% chance that a person has the infection. An infected person has an 80% chance of testing positive. What are the chances that a person has the infection and tests positive? <u>[4]</u> %
Natural frequency	There are 5 people out of 100 who have the infection. Among the infected people, 4 out of 5 test positive. How many people have the infection and test positive? <u>[4]</u> out of 100
Normalized frequency	There are 5% of people who have the infection. Among the infected people, 80% test positive. How many people have the infection and test positive? <u>[4]</u> %

degree (62%) or high school diploma (37%), and only a few of them (1%) had a lower educational level. The participants were paid \$1 and were randomly assigned to one of the four groups ($n = 45$ each).

Materials and design Participants had to interpret one statement by indicating the type of information that, according to them, it conveyed. They did so by selecting either the frequency interpretation (i.e., “The frequency of people who have the infection and test positive”) or the single-event probability interpretation (i.e., “The probability that a person has the infection and tests positive”). The presentation order of the two alternative responses was randomized. After the interpretation task, participants were asked to answer a reasoning question concerning the sensitivity of the test (i.e., the true positive rate; see Table 1).

Results

Table 2 reports the numbers of single-event probability interpretations for the four statements, as well as the corresponding statistical results according to a binomial test. In line with Girotto and Gonzalez’s (2001) theoretical argument, the majority of participants gave a single-event probability interpretation of the unnormalized chance statement, while for the natural frequency statement the opposite was the case (78% vs. 36% single-event probability interpretations, respectively) [$\chi^2(1) = 16.3, p < .001, w = .43$]. A similar pattern was found for the normalized chance and normalized frequency statements (76% vs. 31% single-event probability interpretations, respectively) [$\chi^2(1) = 17.8, p < .001, w = .44$]. Participants’ interpretations of the unnormalized and normalized chance statements did not differ significantly [$\chi^2(1) = 0.06, p = .803$], nor did their interpretations of the natural and normalized frequency statements [$\chi^2(1) = 0.20, p = .655$]. To quantify the evidence in support of the null hypothesis in the last two comparisons, we computed the corresponding Bayes factors (using JASP 0.7.5.6; www.jasp-stats.org). The analysis (assuming a uniform distribution of priors; see Albert, 2009) yielded anecdotal

evidence for the null hypothesis that participants’ categorizations did not differ either between unnormalized and normalized chances ($BF_{01} = 2.89$) or between natural and normalized frequencies ($BF_{01} = 2.51$). These results suggest that laypeople are able to discriminate between chance and frequency statements independently from the normalization of the numerical information.

Four separate logistic regression analyses showed that participants’ interpretations of the numerical formats made no significant contribution to predicting their accuracy in the reasoning question³ (the Wald criteria are reported in Table 2). To further quantify the support in favor of the hypothesis that interpretation had a significant effect on participants’ accuracy, we calculated the corresponding Bayes factors.⁴ As we report in Table 2, this analysis yielded only anecdotal (for unnormalized chance and normalized frequency statements) or substantial (for normalized chance and natural frequency statements) evidence that participants’ interpretations affected their reasoning accuracy rates. Note that this (weak) support was in a direction somewhat opposite to Brase’s (2008)

³ To investigate whether education level influenced participants’ reasoning accuracy and/or interpretations, we included education as a categorical covariate in the logistic regressions. Education level did not influence participants’ interpretations, and it was a significant predictor of accuracy only in the normalized-frequency group [$\chi^2(1) = 6.12, p = .013$]. Note, however, that the participants in Experiment 1 were not uniformly distributed across the three educational levels considered.

⁴ The Bayes factor was computed as follows:

$$BF_{01} = \frac{BIN(n_{I_{p,c}}; n_{I_p}; p(c|H_0))BIN(n_{I_{f,c}}; n_{I_f}; p(c|H_0))}{BIN(n_{I_{p,c}}; n_{I_p}; p(c|H_{1_p}))BIN(n_{I_{f,c}}; n_{I_f}; p(c|H_{1_f}))}$$

where n_{I_p} (n_{I_f}) is the number of single-event probability (or frequency) interpretations; $n_{I_{p,c}}$ ($n_{I_{p,\sim c}}$) is the number of correct (or incorrect) responses associated with a single-event probability interpretation; $n_{I_{f,c}}$ ($n_{I_{f,\sim c}}$) is the number of correct (or incorrect) responses associated with a frequency interpretation, and the probabilities associated with the binomial distributions are computed as follows:

$$p(c|H_0) = \frac{n_{I_{p,c}} + n_{I_{f,c}}}{N}, p(c|H_{1_p}) = \frac{n_{I_{p,c}}}{n_{I_p}}, p(c|H_{1_f}) = \frac{n_{I_{f,c}}}{n_{I_f}}$$

Table 2 Experiment 1 results

Numerical Format	Single-Event Probability Interpretation	Binomial Test p Value and Corresponding Effect Size	Conditional Correct Responses			
			Single-Event Probability Interpretation	Frequency Interpretation	Wald Criterion	BF_{01}
Unnormalized chance	35 (78%)	$<.001, \rho = .35$	26/35 (74%)	7/10 (70%)	$\chi^2(1) = 0.07, p = .79$	0.9642
Normalized chance	34 (76%)	$<.001, \rho = .32$	17/34 (50%)	2/11 (18%)	$\chi^2(1) = 3.10, p = .08$	0.1552
Natural frequency	16 (36%)	$=.072, \rho = .17$	14/16 (87%)	18/29 (62%)	$\chi^2(1) = 2.94, p = .09$	0.1796
Normalized frequency	14 (31%)	$<.05, \rho = .23$	3/16 (21%)	11/31 (35%)	$\chi^2(1) = 0.87, p = .35$	0.6306

Reported in the second column from the left are the occurrences (and the corresponding percentages) of the single-event probability interpretation for the four statements employed in Experiment 1 (the frequency interpretation is complementary). The third column contains the results of the binomial test. Columns 4 and 5 report the occurrences (and the corresponding percentages) of correct responses to the reasoning question conditional on (single-event probability vs. frequency) interpretation. The last two columns report the results of the logistic regressions and the corresponding Bayes factors, which indicate the contribution of participants' interpretations to their reasoning accuracy.

conclusion: Participants' reasoning performance tended to be slightly more accurate for those who gave a single-event probability interpretation to chance statements (and even to the natural frequency statement).

In sum, our results indicate that both the unnormalized and normalized chance statements were interpreted by the great majority of participants as conveying single-event probability information, and that this was unrelated to any decrease in their reasoning performance. On the other hand, participants' interpretations of natural frequencies and the effects of those interpretations on reasoning performance were slightly more ambiguous.

Experiment 2

The aim of Experiment 2 was to further control, using a more traditional Bayesian inference problem, whether the facilitation that can be obtained with chances depends on their being interpreted as natural frequencies. In order to have an adequate sample size of laypeople who provided the correct solution, we employed both the two-step and distributive questions (Giroto & Gonzalez, 2001; Pighin, Gonzalez, Savadori, & Giroto, 2015), which are typically associated with good accuracy rates. The former asks for separate computations of the two terms of a ratio, forcing individuals to consider both relevant values, and the latter asks for the numbers of chances favoring one hypothesis against its alternative.

Method

Participants According to a power analysis with $\alpha = .05$, $1 - \beta = .95$, and a medium effect size ($w = .4$), a minimum

sample size of 82 participants per group was required. Since we wanted to have four groups (see below), a new sample of 336 U.S. residents ($M_{\text{age}} = 38$ years, $SD = 13.1$; 173 men, 163 women) was recruited through the AMT platform. Participants received a payment of \$1 on completion of the task. Most of them had a university degree (56%) or had graduated high school (42%), and only a few participants (2%) had a lower educational level.

Materials and design Participants were randomly assigned to one of four groups in a 2 (two-step vs. distributive question) $\times$ 2 (Gold vs. Infection scenario) between-subjects design (the questions and scenarios are reported in Table 3). They were asked first to solve a Bayesian inference problem and then to answer a follow-up interpretive question by indicating whether the information conveyed in the scenario referred to either a single-event probability or a frequency.

Results

In line with previous studies (Brase, 2009; Giroto & Gonzalez, 2001; Pighin, Gonzalez, Savadori, & Giroto, 2015), the accuracy rates for the two-step questions were 32% and 36%, while those for the distributive questions were 62% and 64%, for the gold and infection scenarios, respectively (see Table 4).

In all four groups, the overall number of participants who reported themselves as thinking about a single-event probability while reading and answering the question was significantly higher than 50% (all $ps < .001$, binomial tests; see Table 4).

Two separate chi-square comparisons indicated that the accuracy rates for both the distributive and two-step questions did not differ significantly between the two problems

Table 3 Scenarios and questions (and the corresponding correct responses, within square brackets) employed in Experiment 2

Scenario	
Gold	<p>Pat's hobby is prospecting for gold. For this purpose, Pat uses a gold detector machine, which is a metal detector tuned so that it selectively detects gold by emitting a loud sound.</p> <p>Previous observations on the area that Pat is exploring indicate that if one finds a coin, there are 8 chances out of 1,000 that it is a gold coin. The gold detector machine is not perfect. Specifically, when a gold coin is found, there are 7 chances out of 8 that the gold detector machine correctly emits a loud sound. On the other hand, the gold detector machine emits a loud sound also in 2 of the remaining 992 chances of finding a nongold coin.</p> <p><i>[Two-step question]</i> Pat has found a new coin. (Please fill in the following sentence by entering a single whole number in each blank space) Out of a total of 1,000 chances, there are <u> 9 </u> chance(s) that the gold detector machine emits a loud sound for this new coin, <u> 7 </u> of which are associated with the coin actually being gold.</p> <p><i>[Distributive question]</i> Pat has found a new coin. The gold detector machine emits a loud sound. (Please fill in the following sentence by entering a single whole number in each blank space) For this new coin, there are <u> 7 </u> chances that the gold detector machine emitted a loud sound and the coin is made of gold versus <u> 2 </u> chances that the gold detector machine emitted a loud sound but the coin is not made of gold.</p> <p><i>[Interpretive question]</i> When I read the previous description and answered the question, I thought about: ☐ the frequency of gold coins among those for which the detector machine emits a loud sound. ☐ the probability of finding a gold coin when the detector machine emits a loud sound.</p>
Infection	<p>There is a newly discovered infection and a test that is useful for detecting it. There follows some information about the infection and the test. A person has 6 chances out of 100 of having the infection.</p> <p>The test to detect whether or not a person has the infection is not perfect. Specifically, only 4 chances of the 6 of having the infection are correctly associated with a positive reaction from the test. On the other hand, also 16 chances of the remaining 94 of not having the infection are incorrectly associated with a positive reaction from the test.</p> <p><i>[Two-step question]</i> Imagine that a new person, Pat, undergoes the test. (Please fill in the following sentence by entering a single whole number in each blank space) Out of a total of 100 chances, Pat has <u> 20 </u> chance(s) of positive reaction from the test, <u> 4 </u> of which are associated with actually having the infection.</p> <p><i>[Distributive question]</i> Imagine that a new person, Pat, undergoes the test. Pat has a positive reaction from the test. (Please fill in the following sentence by entering a single whole number in each blank space) For Pat, there are <u> 4 </u> chances that s/he had a positive reaction from the test and actually has the infection versus <u> 16 </u> chances that s/he had a positive reaction from the test but does not have the infection.</p> <p><i>[Interpretive question]</i> When I read the previous description and answered the question, I thought about: ☐ the frequency of infected people among those who have a positive test result. ☐ the probability that a person is infected if s/he has a positive test result.</p>

$[\chi^2(1) = 0.10, p = .749$, and $\chi^2(1) = 0.24, p = .625$, respectively]. These results were confirmed by a Bayes factor analysis, which yielded substantial evidence for the null hypothesis that participants' accuracy did not differ between the two problems, both for the distributive ($BF_{01} = 3.46$) and for the two-step ($BF_{01} = 3.3$) questions. Nor was a significant difference obtained in the follow-up questions among the four experimental conditions [$\chi^2(3) = 1.76, p = .624$; $BF_{01} = 37$]. Accordingly, the responses obtained in the two scenarios were analyzed together.

Two logistic regressions showed that interpreting chance information in terms of either single-event probability or frequency made no significant contribution to predicting the accuracy of either the two-step or the distributive questions

(see Table 4).⁵ As in Experiment 1, to provide an adequate assessment of the extent to which the data collected supported the alternative or the null hypothesis, we also computed the Bayes factors. The results are reported in Table 4 and show only anecdotal evidence in support of the hypothesis that different interpretations had an effect on participants' accuracy rate (in the same direction as in the previous experiment).

In conclusion, in line with the results of Experiment 1, numbers of chances were mostly interpreted as single-event

⁵ Including education level as a covariate in the two logistic regression analyses did not affect the results. Education level did not significantly predict the accuracy in the Bayesian inference task, independently of the type of question employed ($ps = .952$ and $.785$ for the two-step and distributive questions, respectively).

Table 4 Experiment 2 results

Question and Scenario	Correct Responses	Single-Event Probability Interpretation	Binomial Test p Value and Effect Size	Conditional Correct Responses			
				Single-Event Probability Interpretation	Frequency Interpretation	Wald Criterion	BF ₀₁
Two-Step							
Gold	27 (32%)	60 (71%)	<.001, ρ = .24	21/60 (35%)	6/24 (25%)	$\chi^2(1) = 1.14, p = .29$	0.5599
Infection	30 (36%)	61 (73%)	<.001, ρ = .26	23/61 (38%)	7/23 (30%)		
Distributive							
Gold	52 (62%)	59 (70%)	<.001, ρ = .23	37/59 (63%)	15/25 (60%)	$\chi^2(1) = 0.10, p = .75$	0.9542
Infection	54 (64%)	66 (79%)	<.001, ρ = .35	41/66 (62%)	13/18 (72%)		

The second column from the left reports the numbers (and the corresponding percentages) of correct responses in the two Bayesian inference problems for each (two-step vs. distributive) question. The third column reports the numbers (and percentages) of participants who gave a single-event probability interpretation of the chances (frequency interpretations are complementary). The fourth column contains the results of the binomial test. The fifth/sixth columns report the numbers (and percentages) of correct answers among participants who provided a single-event probability/frequentist interpretation. The last two columns report results of the logistic regressions and the corresponding Bayes factors, which indicate the contribution of participants' interpretations to their reasoning accuracy.

probabilities, rather than as natural frequencies. Moreover, in contrast with Brase's (2008) conclusions, laypeople's accuracy in Bayesian reasoning tasks was not improved by a frequentist interpretation of chances.

Discussion

The aim of the research reported in this article was twofold: first, to control empirically whether laypeople interpret numbers of chances as single-event probabilities or, instead, confuse them with natural frequencies. Second, to check whether the facilitation effect of chances in Bayesian reasoning depends on them being interpreted as frequencies. The results showed that the large majority of our participants interpreted both normalized and unnormalized chances (but not natural or normalized frequencies) as single-event probabilities, and that their reasoning accuracy was not increased by possible frequentist interpretations of chances. The same pattern of results was obtained with different scenarios and with the interpretive question either preceding (Exp. 1) or following (Exp. 2) the reasoning question. Overall, therefore, our results supported Girotto and Gonzalez's (2002) theoretical argument that chances are not conflated with natural frequencies, and they strongly disconfirmed Brase's (2008) conclusion that chances lead to an improvement in Bayesian reasoning mainly when they are interpreted as frequencies.

The evidence that laypeople can make sound Bayesian inferences when dealing with genuine single-event probabilities has major theoretical and applied implications.

As regards the theoretical implications, our results are coherent with the nested-sets view, which ascribes the primary source of facilitation in Bayesian reasoning problems to the partitive structure and question forms that prompt individuals

to compute the two terms of the Bayesian ratio. In line with Sirota et al. (2015b), our results also indicate that the reasoning accuracy with chances is not mediated by their frequentist interpretation. Consequently, it appears that there is no need to call upon a cognitively privileged frequentist representational format for Bayesian reasoning. This finding is supported all the more in light of the recent finding (see Pighin, Gonzalez, Savadori, & Girotto, 2016) that the recommendation to present medical test results in terms of natural frequencies is not well-grounded in empirical research and does not necessarily lead to more correct judgments than percentages do.

From an applied perspective, the results obtained represent good news, especially for all those important real-life domains, such as certain kinds of judicial and medical decisions, that elude easy representation in terms of natural frequencies. Consider, for example, a defendant who is interested in the probability of being acquitted if she implements different trial strategies, or a patient with a rare disease who tries to consider her probability of survival if she takes different therapeutic approaches. Natural frequencies may not be available to estimate the probabilities of these events, which refer to a single person. Moreover, this person is likely to consider her specific legal or health condition to be somehow unique and not fully comparable with those of other individuals. Should we therefore conclude that probabilistic inferences in such important situations are not possible? The results of our study suggest that we cannot. Laypeople are able to make sound Bayesian inferences concerning single events, as long as the relevant information is conveyed in a transparent manner and the questions are posed in a way that allows them to reason extensionally by enumerating and counting the possibilities. When single-event probabilities are at stake, far from lacking in Bayesian thinking, the human mind has a real chance to engage in good reasoning.

Author note We are grateful to Phil Johnson-Laird, Donatella Ferrante, and Vincenzo Crupi for their comments on earlier drafts. Financial support for this study was provided by the Cassa di Risparmio di Trento and Rovereto (SMC, Grant 40102595) and by the Deutsche Forschungsgemeinschaft as part of the priority program New Frameworks of Rationality (SPP 1516, Grant CR 409/1-2). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The study was approved by the Ethics Committee of CERME (University Ca' Foscari of Venice). The data are available online at DOI:10.17605/OSF.IO/7KDRC.

References

- Albert, J. (2009). *Bayesian computation with R*. New York: Springer.
- Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30, 241–297.
- Barton, A., Mousavi, S., & Stevens, J. R. (2007). A statistical taxonomy and another “chance” for natural frequencies. *Behavioral and Brain Sciences*, 30, 255–256. doi:10.1017/S0140525X07001665
- Brase, G. L. (2002). Ecological and evolutionary validity: Comments on Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni's (1999) mental-model theory of extensional reasoning. *Psychological Review*, 109, 722–728. doi:10.1037/0033-295X.109.4.722
- Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, 15, 284–289. doi:10.3758/PBR.15.2.284
- Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, 23, 369–381.
- Brase, G. L., & Hill, W. T. (2015). Good fences make for good neighbors but bad science: A review of what improves Bayesian reasoning and why. *Frontiers in Psychology*, 6, 340. doi:10.3389/fpsyg.2015.00340
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Evans, J. S. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, 77, 197–213.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160. doi:10.3758/BRM.41.4.1149
- Fox, C. R., & Levav, J. (2004). Partition–edit–count: Naïve extensional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General*, 133, 626–642. doi:10.1037/0096-3445.133.4.626
- Gigerenzer, G. (1996). The psychology of good judgment: Frequency formats and simple algorithms. *Medical Decision Making*, 16, 273–280.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency format. *Psychological Review*, 102, 684–704. doi:10.1037/0033-295X.102.4.684
- Gigerenzer, G., & Hoffrage, U. (2007). The role of representation in Bayesian reasoning: Correcting common misconceptions. *Behavioral and Brain Sciences*, 30, 264–267. doi:10.1017/S0140525X07001756
- Gigerenzer, G., Hoffrage, U., & Ebert, A. (1998). AIDS counseling for low-risk clients. *AIDS Care*, 10, 197–211.
- Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, 78, 247–276.
- Girotto, V., & Gonzalez, M. (2002). Chances and frequencies in probabilistic reasoning: Rejoinder to Hoffrage, Gigerenzer, Krauss, and Martignon. *Cognition*, 84, 353–359.
- Hill, W. T., & Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *Quarterly Journal of Experimental Psychology*, 65, 2343–2368. doi:10.1080/17470218.2012.687004
- Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, 73, 538–540.
- Hoffrage, U., & Gigerenzer, G. (2004). How to improve the diagnostic inferences of medical experts. In E. Kurz-Milcke & G. Gigerenzer (Eds.), *Experts in science and society* (pp. 249–268). Dordrecht: Kluwer Academic.
- Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, 84, 343–352.
- Hoffrage, U., Krauss, S., Martignon, L., & Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Frontiers in Psychology*, 6, 1473. doi:10.3389/fpsyg.2015.01473
- Johnson, E. D., & Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Frontiers in Psychology*, 6, 938. doi:10.3389/fpsyg.2015.00938
- Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Sonino Legrenzi, M., & Caverni, J. P. (1999). Naïve probability: A mental model theory of extensional reasoning. *Psychological Review*, 106, 62–88. doi:10.1037/0033-295X.106.1.62
- Kynn, M. (2008). The “heuristics and biases” bias in expert elicitation. *Journal of the Royal Statistical Society*, 171, 239–264.
- Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. *Quarterly Journal of Experimental Psychology*, 48, 188–207.
- Moro, R., Bodanza, G. A., & Freidin, E. (2011). Sets or frequencies? How to help people solve conditional probability problems. *Journal of Cognitive Psychology*, 23, 843–857. doi:10.1080/20445911.2011.579072
- Pighin, S., Gonzalez, M., Savadori, L., & Girotto, V. (2015). Improving public interpretation of probabilistic test results: Distributive evaluations. *Medical Decision Making*, 35, 12–15.
- Pighin, S., Gonzalez, M., Savadori, L., & Girotto, V. (2016). Natural frequencies do not foster public understanding of medical test results. *Medical Decision Making*, 36, 686–691.
- Sirota, M., Kostovičová, L., & Juanchich, M. (2014). The effect of iconicity of visual displays on statistical reasoning: Evidence in favor of the null hypothesis. *Psychonomic Bulletin & Review*, 21, 961–968. doi:10.3758/s13423-013-0555-4
- Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015a). How to train your Bayesian: A problem-representation transfer rather than a format-representation shift explains training effects. *Quarterly Journal of Experimental Psychology*, 68, 1–9. doi:10.1080/17470218.2014.972420
- Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015b). Now you Bayes, now you don't: Effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychonomic Bulletin & Review*, 22, 1465–1473. doi:10.3758/s13423-015-0810-y
- Sloman, S. A., Over, D. E., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, 91, 296–309.