From *reading* numbers to *seeing* ratios: A benefit of icons for risk comprehension

Elisabet Tubau[1,2], Javier Rodríguez-Ferreiro[1,2], Itxaso Barberia[1] and Àngels Colomé[1,2]

[1]Department of Cognition, Development and Educational Psychology

[2]Institute of Neurosciences (UBNeuroscience)

Universitat de Barcelona

Running Head: Seeing ratios

Corresponding author:

Elisabet Tubau
Departament de Cognició, Desenvolupament i Psicologia de l'Educació
Facultat de Psicologia
Universitat de Barcelona
Passeig de la Vall d'Hebron, 171
08035 Barcelona
Catalonia-Spain

Email: etubau@ub.edu

Abstract

Promoting a better understanding of statistical data is becoming increasingly important for improving risk comprehension and decision-making. In this regard, previous studies on Bayesian problem solving have shown that iconic representations help infer frequencies in sets and subsets. Nevertheless, the mechanisms by which icons enhance performance remain unclear. Here, we tested the hypothesis that the benefit offered by icon arrays lies in a better alignment between presented and requested relationships, which should facilitate the comprehension of the requested ratio beyond the represented quantities. To this end, we analyzed individual risk estimates based on data presented either in standard verbal presentations (percentages and natural frequency formats) or as icon arrays. Compared to the other formats, icons led to estimates that were more accurate and, importantly, promoted the use of equivalent expressions for the requested probability. Furthermore, whereas the accuracy of the estimates based on verbal formats depended on their alignment with the text, all the estimates based on icons were equally accurate. Therefore, these results support the proposal that icons enhance the comprehension of the ratio and its mapping onto the requested probability and point to relational misalignment as potential interference for text-based Bayesian reasoning. The present findings also argue against an intrinsic difficulty with understanding single-event probabilities.

Keywords: probabilistic reasoning, iconic representation, relational alignment, Bayesian problem solving, risk comprehension

Introduction

Everyday decision-making, in areas ranging from healthcare to finance, often requires the integration of different pieces of statistical information to infer the probability of relevant outcomes. For example, the decision to participate in breast cancer screening (i.e., mammograms) depends, among other factors, on the perceived predictive value of the test (e.g., Navarrete et al., 2015). This commonly requires considering different data: the breast cancer prevalence (the base rate), and the conditional probabilities of a positive mammogram in the presence (hit rate) and in the absence (false-alarm rate) of breast cancer. The cognitive demands involved in this inference (comprehension of the data and the corresponding arithmetic calculations), as well as potentially helpful strategies, have been the subject of widespread research in the field of Bayesian reasoning (e.g., see recent reviews in Mandel & Navarrete, 2015).

In a typical Bayesian problem, solvers are presented with the above information (base rate, hit rate and false-alarm rate) and required to calculate the Bayesian probability (e.g., the posterior probability of having breast cancer in the case of receiving a positive mammogram; see examples in the appendix). It is well known that this is no trivial task; the percentage of participants producing accurate estimates is often lower than 40%, even for problems in which the data are presented in natural frequency format (frequencies that preserve the reference class as “3 of 4” instead of normalized ratios such as “75%”; e.g., Barbey & Sloman, 2007; Brase, 2009; Chapman & Liu, 2009; Evans et al., 2000; Gigerenzer, 1991; Gigerenzer & Hoffrage, 1995; Hoffrage et al., 2015; Pighin et al, 2016; Sloman et al., 2003; Sirota et al., 2014a,b; Sloman et al., 2003; see also the meta-analysis by McDowell & Jacobs, 2017).

As suggested recently by Johnson & Tubau (2017), Bayesian word-problems reporting natural frequencies can be as difficult as the ones reporting percentages

because they do not eliminate the relational misalignment between presented and requested relations. More specifically, in standard verbal presentations, numbers are associated with either the set or the subset of the relationship. For instance, numbers in bold in the examples "of 4 women with breast cancer, **3 receive a positive mammogram**" and "of 96 women without breast cancer, **12 receive a positive mammogram"** specify the size of corresponding subset. As observed in similarity judgments and analogical reasoning, "objects are placed in correspondence based on their roles within the matching relational structure" (Markman & Gentner, 1993, p. 459). Accordingly, number-relational role associations induced from standard presentations make it possible to use such numbers in a similar role ("of 4+96 women, **3+12 receive a positive mammogram**"), but hinder their use in a different role, as required in the Bayesian inference "of **3+12 women with positive mammogram**, 3 have breast cancer," where the bold part indicates the set (posterior reference class in Figure 1). In this regard, inaccurate Bayesian reasoning might be caused not only by a limited understanding of the nested-set structure of the data (Barbey & Sloman, 2007), or how to translate frequencies into probabilities (Cosmides & Tooby, 1996; Gigerenzer, 1991), but by difficulties involved in mapping the presented relationships onto the requested one (preliminary evidences supporting this proposal, using natural frequencies, are reported in Johnson & Tubau, 2017).

Figure 1

Interestingly, and also consistent with this proposal, problems that present the sample statistics as frequency grids or icon arrays make it easier to solve the Bayesian question in frequency format, compared to verbal presentations alone (e.g., Brase, 2009 and 2014; Galesic, Garcia-Retamero & Gigerenzer, 2009; Garcia-Retamero & Hoffrage, 2013; Sedlmeier & Gigerenzer, 2001; but see Brase & Hill, 2017; Cosmides & Tooby,

1996 and Sirota et al, 2014b for mixed findings). Critically, by explicitly presenting the requested set and subset in overlapping areas, icons reduce the relational reasoning demand of the Bayesian inference (i.e., the posterior ratio can be *seen at a glance*; see Figure 2). The benefit of icons seems to be stronger when they are presented without the redundant text (e.g., Khan et al., 2015; Ottley et al., 2016), a fact that suggests that standard verbal presentations promote the formation of misleading associations (Barbey & Sloman, 2007; Johnson & Tubau, 2015). Indeed, the null benefit of icons (see references above) was observed in icons+text presentations. Hence, a better relational alignment, together with reduced interference from misleading verbal associations, might explain the benefit of icons for Bayesian problem solving.

Figure 2

Previous studies have demonstrated the benefit of icons for inferring either frequencies in subsets or individual chances, but in both cases through questions that refer to the specific quantities represented in the array (e.g., "Imagine Michael is tested now. Out of a total of 100 chances, Michael has _____ chance(s) of a positive reaction from the test, _____ of which will be associated with actually having the infection"; Brase, 2009). However, a stronger proof of the usefulness of icons for understanding individual risks would be provided by requesting estimates of individual probabilities, without prompting a determined reference class. By using more ambiguous questions, it would be possible to study the extent to which icons enhance comprehension of the ratio, beyond the represented quantities. If this were the case, icons might induce more accurate estimates and the use of equivalent expressions; that is, the use of different numbers to express the same probability (e.g., the posterior probability in the mammogram problem presented in the appendix can be expressed as "3 of 15", "1 of 5", "20 of 100" or "2 of 10"). The present research aimed to test this hypothesis by

analyzing the form and the accuracy of participants' probability estimates, based on data presented either in iconic or verbal formats. Given that natural frequency problems commonly prompt to infer frequencies in determined set and subset (frequency question), it is uncertain the extent to which, compared to percentages, natural frequencies facilitate inferring single-event probabilities. In this sense, a second goal of this research was to shed light into this issue by using a probability question. Finally, it also aimed to test whether the differences between formats might depend on the alignment of the request (aligned or misaligned with the presented relationships).

Experiment 1

To test the benefit of icons for ratio comprehension, problems that presented the sample statistics in one of three formats (icon arrays: IA; natural frequencies: NF; and percentages: PE) and that requested a single-event probability were presented to three groups (see the appendix). In contrast to PE problems, which unambiguously prompt the use of a percentage, we expected to observe different interpretations of the requested subset and reference class in the responses to IA and NF problems. Due to the increased computational demands, we also expected less accurate normalized responses (e.g., percentages) than non-normalized ones (e.g., ratios of represented frequencies). Nevertheless, if icons enhance comprehension of the requested ratio, compared to the verbal formats, they should facilitate its expression in any equivalent form.

Method

Participants

One hundred and forty (36 men and 104 women; mean age=22.87 years, SD=5.57) psychology undergraduates from the University of Barcelona took part in this experiment before being introduced to Bayes' rule. All of them provided written consent and the research was approved by the University of Barcelona's Bioethics

Commission. Participants were randomly assigned to three groups according to the format in which the data were presented (icon arrays: N=49; natural frequencies: N= 49; percentages: N=42). Given that each participant solved two problems (see below), we analyzed more than 80 responses for each condition.

Materials and procedure

All participants had to evaluate the two health scenarios shown in the appendix, with the data presented in one of the three formats: IA, NF or PE. The single-event probability question was identical across the three groups. Participants were tested collectively, but each had their own computer and solved the task individually by typing the requested responses in "X of Y" or "%" in PE format (see question in the appendix). There were no time limits for responses, but all participants finished the whole exercise in less than 20 minutes.

Results and discussion

Responses were coded as correct when the division of the proposed numbers matched the mathematical probability, i.e., 0.2 in the mammogram problem and 0.33 in the hypertension problem. For the latter, responses rounded to 0.3 were also considered correct (IA: 5 responses; NF: 1 response; PE: 4 responses). Given the match between this rounded response and the false-alarm rate, rounded responses expressed as "24 of 80" were not counted as correct (1 response in NF format[1]). Accuracy levels for the two scenarios were similar in each format (*ps*>.11), so the analyses were performed by taking the total of correct responses (0, 1 or 2) for each participant into account.

As expected, we observed that the problem format had a significant effect on accuracy ($\chi^2(4)=89.16$, $p<.001$, $V=.56$; see Figure 3). The mean numbers of correct

[1] We are aware that the response "30%" also coincides with the literal representation of the false-alarm rate for the PE group. Nevertheless, given the extremely low percent accuracy for this group (0% according to the strict criterion; 5% considering the rounded responses), the reported effects are independent of the interpretation of this ambiguity.

estimates for IA, NF and PE groups were 1.54, 0.22 and 0.1, respectively. Differences were significant between IA and NF groups ($\chi^2(2)=55.82$, $p<.001$, $V=.75$), and between IA and PE groups ($\chi^2(2)=61.15$, $p<.001$, $V=.82$). No significant difference was found between NF and PE groups ($\chi^2(2)=2.41$, $p=.30$, $V=16$).

Figure 3

Regarding the response format, responses to NF problems included the 100 in the denominator more often than responses to IA (Mammogram scenario: 63% vs 31%; $\chi2(1)=10.49$, $p=.001$, $\phi=.33$; Hypertension scenario: 71% vs. 33%; $\chi2(1)=14.76$, $p<.001$, $\phi=.39$). As expected, accuracy was lower for these apparently normalized responses[2] than for non-normalized responses in each format and for each problem (IA-mammogram: $\chi^2(1)=11.43$, $p=.001$, $\phi=.49$; IA-hypertension: $\chi^2(1)=5.54$, $p=.02$, $\phi=.38$; NF-mammogram: $\chi^2(1)=11.38$, $p=.001$, $\phi=.49$), except for the NF-hypertension scenario ($p>.21$), due to the very few correct responses (see Table 1). Crucially, responses to IA problems were more accurate than responses to NF problems, either among normalized (Mammogram scenario: $\chi^2(1)=13.28$, $p<.001$, $\phi=.54$; Hypertension scenario: $\chi^2(1)=23.1$, $p<.001$, $\phi=.67$) or non-normalized responses (Mammogram scenario: $\chi^2(1)=14.42$, $p<.001$, $\phi=.54$; Hypertension scenario: $\chi^2(1)=25.83$, $p<.001$, $\phi=.75$). Moreover, as shown in Table 1, 34% of the correct non-normalized responses to IA problems (in total 20 of 59) were simplified ratios (e.g., “1 of 3” instead of “12 of 36”), whereas a single correct response to NF problems was expressed as a simplified ratio.

[2] We considered the use of 100 or 10 (only observed in IA responses) in the denominator as an attempt to normalize the response. Nevertheless, as discussed below, most of the NF responses using the 100 might not be “true” normalization attempts, but rather a consequence of misleading associations.

Table 1. Overall percentage of correct responses, and corresponding frequencies for each scenario, among normalized (including 100 or 10 as denominator), and non-normalized ratios in Experiment 1 (frequencies of correct simplified ratios, as "1/3" instead of "12/36", are shown within parentheses).

	Type of ratio			
	Normalized		**Non-Normalized**	
Icon Array[1]	55%	Mammogram: 7 of 15	91%	Mammogram: 30 of 33 (7)
		Hypertension: 10 of 16		Hypertension: 29 of 32 (13)
Natural Frequencies[2]	3%	Mammogram: 1 of 31	29%	Mammogram: 7 of 17 (1)
		Hypertension: 1 of 35		Hypertension: 2 of 14 (0)
Percentages	5%	Mammogram: 0 of 42		
		Hypertension: 4 of 42		

[1] Two participants solved only one of the problems. [2] One participant solved only the Hypertension problem.

An analysis of the errors showed differences between the scenarios (see Figure 4). For the mammogram scenario, the most common errors were caused by confusion with the hit rate (12% and 36% for NF and PE formats, respectively), the total positive rate (22% and 10% for NF and PE formats, respectively) and the base rate (10% for either NF or PE format). For the hypertension scenario, the most common sources of confusion were the base rate (30% and 29% for NF and PE formats, respectively) and the hit rate (13% and 10%, for NF and PE formats, respectively). The few errors detected in the IA responses were equally distributed among the abovementioned categories.

Figure 4

In sum, the probability estimates based on icons were more accurate and expressed in more diverse equivalent forms than those based on verbal formats. These findings support the hypothesis that icons facilitate the comprehension of the ratio beyond the represented quantities. Furthermore, the large distribution of errors in both verbal formats confirmed the suggestion that verbal presentations induce superficial reasoning and misleading associations (Barbey & Sloman, 2007; Johnson & Tubau, 2017). Results also suggest that natural frequencies, like percentages, are unhelpful for inferring single-

event probabilities. Nevertheless, as previously shown in the context of frequency estimates (Johnson & Tubau, 2017), this limitation might be related to the misalignment of the Bayesian inference. Experiment 2 aimed to test this hypothesis.

Experiment 2

Based on the alignment hypothesis (Johnson & Tubau, 2017), we hypothesized that icons would facilitate the comprehension of the ratio through a more direct mapping between the data and the request. Nevertheless, alternative explanations might also hold. The advantage of icons has been attributed to their role enhancing a frequentist interpretation of the requested chances (Brase, 2009 and 2014; Cosmides & Tooby, 1996), or a clearer representation of the nested-set structure of the data (Barbey & Sloman, 2007; Reyna, 2004). Therefore, besides theoretical discrepancies between these accounts (see for example the comments on Barbey & Sloman, 2007), both would predict differences between iconic and verbal formats for estimates requiring identical computation. In contrast, the relational alignment hypothesis would predict differences between formats mainly in case of misalignment between presented and requested relationships.

In order to test these hypotheses, new groups of participants saw/read the previous IA or NF data, but were requested two single-event probability estimates requiring the same arithmetical steps but differing in their alignment with the text (see Materials and procedure). Based on the frequentist or nested-sets accounts, we expected a significant effect of format for both estimates. Nevertheless, from the relational alignment hypothesis, we expected differences between formats mainly for the misaligned estimate.

Participants

One hundred and sixteen students (16 men and 100 women; mean age=21.28 years, SD=2.68) from the same population as in Experiment 1 took part in this experiment. All of them also provided written consent. Participants were randomly assigned to two groups according to the format in which the data were presented (icon arrays: N=57; natural frequencies: N= 59). None of them had participated in similar experiments before.

Materials and procedure

As in the previous experiment, all the participants had to evaluate the two health scenarios shown in the appendix, with the data presented in one of two formats: IA or NF. The problems ended with two single-event probability requests: the aligned one required to estimate the probability of the datum (e.g., *what is the probability of a woman at that age to get a positive mammogram?*), whereas the misaligned one required to estimate the posterior probability of suffering the disease, knowing the datum (see the appendix). Note that both responses require adding the same subsets but for a different role: as a new subset "(3+12) of 100" in the aligned response, or as a new reference class "3 of (3+12)" in the misaligned one. Furthermore, as shown in the appendix, we changed the numbers in the hypertension scenario to avoid the coincidence between the false positive rate and the rounded Bayesian estimate.

Results and discussion

Responses were coded as correct when the division of the proposed numbers matched the mathematical probability (0.15 and 0.24 for the aligned estimate; 0.2 and 0.33 for the misaligned one for mammogram and hypertension scenarios, respectively). For the hypertension scenario, the posterior probability estimate rounded to 0.3 was also considered correct (IA: 4 responses; NF: 3 responses). Accuracy levels for the two

scenarios were similar in each format (*ps*>.48), so the analyses were performed with the total of correct responses (0, 1 or 2) for each participant and question (aligned and misaligned). Figure 5 shows the percentage of participants in each category.

Figure 5

For the aligned requests, the mean number of correct estimates was similar in both groups: 1.44 and 1.38 for the IA and the NF groups, respectively (p=.25). For the misaligned ones, results replicated previous findings: the mean number of correct estimates was higher for the IA group than for the NF group (1.36 vs 0.44; $\chi^2(2)=32.64$, $p<.001$, V=.53; see Figure 5), and the difference between groups was significant for either normalized (Mammogram scenario: $\chi^2(1)=21.32$, $p<.001$, $\phi=.50$; Hypertension scenario: $\chi^2(1)=19.10$, $p<.001$, $\phi=.48$), or non-normalized responses (Mammogram scenario: $\chi^2(1)=4.08$, $p=.04$, $\phi=.36$; Hypertension scenario: $\chi^2(1)=5.26$, $p=.02$, $\phi=.41$). Correct simplifications of the ratio were also more common among responses to IA problems (10 responses in total[3]) than to NF problems (2 responses; see Table 2). Hence, these findings supported the hypothesis that natural frequencies verbally presented would be particularly misleading for misaligned requests, being as useful as icons for inferring single-event probabilities from aligned relationships. In contrast, by avoiding the directionality of the text, icons were equally helpful for any probability estimate.

[3] In Experiment 2, most of the Bayesian (misaligned) ratios used 10 or 100 as denominator (66% and 80% for IA and NF formats, respectively). Therefore, although simplifications were fewer than in Experiment 1, the overall percentage of correct responses to IA problems expressed as equivalent ratios was indeed higher (49% vs 67% for Experiments 1 vs 2).

Table 2. Overall percentage of correct responses to the misaligned question, and corresponding frequencies for each scenario, among normalized (including 100 or 10 as denominator) and non-normalized ratios in Experiment 2 (frequencies of correct simplified ratios, as "1/3" instead of "8/24", are shown within parentheses).

	Type of ratio			
	Normalized		**Non-Normalized**	
Icon Array[1]	56%	Mammogram: 23 of 38	95%	Mammogram: 17 of 18 (3)
		Hypertension: 19 of 37		Hypertension: 19 of 20 (7)
Natural Frequencies	11%	Mammogram: 6 of 47	67%	Mammogram: 7 of 12 (1)
		Hypertension: 4 of 47		Hypertension: 9 of 12 (1)

[1] One participant solved only the Hypertension problem.

General discussion

The present research aimed to test the hypothesis that icon arrays facilitate Bayesian reasoning by enhancing comprehension of the ratio, beyond the represented quantities. This proposal was supported by the results of both experiments, which showed that icon arrays not only promoted selection of the correct numerator and denominator, but also induced further numerical processing, as demonstrated by the use of correct equivalent expressions for the requested probability. Results of Experiment 2 also showed that icons promoted equally accurate estimates for any request (probability of the datum or posterior probability). In contrast, for natural frequencies, whereas estimates aligned with the text (probability of the datum) were as accurate as the ones based on icons, the misaligned estimates (posterior probability) were mostly inaccurate. Therefore, a critical difficulty for Bayesian reasoning based on verbal formats (including either percentages or natural frequencies) seems to be the misalignment between presented and requested relationships (Johnson & Tubau, 2017). This relational misalignment might explain the dependency of the Bayesian inference on capacities and skills beyond numeracy (e.g., Chapman & Liu, 2001), such as working memory or reflective thinking (e.g., Lessage et al, 2013; Sirota et al, 2014a).

Importantly, the single-event posterior probability estimates based on icons were as accurate as frequency estimates, as shown in a pilot experiment[4]. This finding argues against an intrinsic difficulty with understanding single-event probabilities (e.g., Cosmides & Tooby, 1996; see similar claims in Girotto & Gonzalez, 2001; Johnson-Laird et al., 1999; Pighin et al., 2017). Specifically, the tendency to simplify the description of the sample statistics (e.g., from "3 of 15" to "1 of 5" or "20%"), observed in half (Experiment 1) or in most (Experiment 2) of the correct responses to IA problems, might point towards the conceptualization of probability as an individual propensity (Gillies, 2000), or as a subjective degree of confidence induced from the sample statistics (e.g., Cosmides & Tooby, 1996[5]). Accordingly, icons might promote a *gist* comprehension of the risk; that is, a comprehension of the numerical relation beyond the specific numbers (e.g., Reyna, 2004). Equivalent expressions were less frequent among the responses to NF problems, which would suggest that verbal presentations promote a more superficial processing of the data.

Of note was the finding that, among correct estimates of the posterior ratio in Experiment 2, normalizations were more frequent than in Experiment 1 in both formats (NF: 2 vs 10; IA: 17 vs 42, for experiments 1 vs 2). This might stem from the influence of the first request, which prompted to use the total sample of 100 as reference class. Nevertheless, the posterior probability (misaligned) estimates based on natural

[4] In the pilot experiment, different groups received the same IA and NF problems of Experiment 1 (see appendix), but were asked for frequencies (e.g., "of the women who test positive, *how many* have breast cancer?"), in the IA (N=20) and NF (N=22) formats. For the IA group, the mean number of correct responses was similar as in the present experiments (1.42). For the NF group, it was higher than in present experiments (0.82), but still lower than for the IA group (p=.02).

[5] Although a default frequency-based representation is defended by these authors, it is also claimed that a frequentist mechanism might produce subjective confidence for single event probabilities: "even though it might initially output a frequency, and perhaps even store the information as such, other mechanisms may make that frequentist output consciously accessible in the form of a subjective degree of confidence" (Cosmides & Tooby, 1996, p. 66, note 19).

frequencies were still mostly inaccurate, even for participants who produced accurate probability (aligned) estimates of the datum[6].

It is also worth noting that, in line with previous observations (e.g., Evans et al., 2000; Hafenbrändl & Hoffrage, 2015; Johnson & Tubau, 2017; Pennycook & Thompson, 2012), the analysis of incorrect responses to either NF or PE problems of Experiment 1 showed highly variable estimates of the posterior probability (between .03 and .95; see Figure 4). This is also coherent with superficial processing of relevant numerals, without the necessary integration, which might be caused by the relational misalignment between presented and requested set-subset relationships (see also Holyoak & Koh, 1987, for similar arguments in other problem-solving tasks). Differences in the frequency of specific errors also confirm the influence of superficial traits as the numerical format (the hit rate was more often selected when presented as a percentage; see also Hafenbrändl & Hoffrage, 2015), the relative magnitude of presented numbers (the base rate .2 in the hypertension scenario was selected more often than the base rate .04 in the mammogram scenario), or the ease of performing certain arithmetic calculations (e.g., the total positive 3+12 in the mammogram scenario was selected more often than the total high sodium diet 12+24 in the hypertension scenario; see Figure 4). The responses to IA problems were affected by these superficial traits to a much lesser degree, with only a few attributed to the abovementioned errors. Accordingly, more integrated pictures such as icon arrays might be useful tools for overcoming *misleading* "associative tendencies" (Barbey & Sloman, 2007).

In conclusion, an important step towards facilitating probabilistic reasoning consists of enhancing the comprehension of statistical data and the corresponding mapping onto the required estimate. In that regard, the present findings confirm the

[6] For NF problems, the mean number of correct posterior probability estimates, among participants who correctly estimated both probabilities of the datum, was 0.5. For IA problems, this mean was 1.6.

advantage of icon arrays for fostering a visual grasp of quantitative relationships. Importantly, as demonstrated by the use of equivalent expressions for the requested probability, present findings suggest that icons enhance the comprehension of the ratio beyond the represented frequencies. Therefore, a much better understanding of individual risks can be achieved by promoting the apprehension of ratios rather than numbers. Whether icons would enhance single-event probabilistic reasoning in other non-university population, or to what extent they would benefit actual decision making, are some of the remaining questions that require further research.

Compliance with Ethical Standards

Funding: ET was funded by Secretaría de Estado de Investigación, Desarrollo e Innovación of Spanish Ministerio de Economía y Competividad (PSI2017-83493-R). AC and ET were also funded by the Catalan Government (2017SGR-48).

Conflict of Interest: All authors declare that they have no conflict of interest.

Ethical approval: Experimental procedure was in accordance with the ethical standards of the University of Barcelona's Bioethics Commission, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

Barbey, A. K. & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30, 241–297.

Brase, G.L. (2009). Pictorial representations and numerical representations in Bayesian reasoning. *Applied Cognitive Psychology*, 23(3), 369–381.

Brase, G.L. (2014). The power of representation and interpretation: Doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *Journal of Cognitive Psychology*, 26, 81–97.

Brase, G. L., & Hill, W. T. (2017). Adding up to good Bayesian reasoning: Problem format manipulations and individual skill differences. *Journal of Experimental Psychology: General*, *146*(4), 577.

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

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all?: Rethinking some conclusions of the literature on judgment under uncertainty. *Cognition*, 58, 1–73.

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(3), 197–213.

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychology*, *28*(2), 210.

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.

Garcia-Retamero, R., Cokely, E.T. & Hoffrage, U. (2015). Visual Aids Improve Diagnostic Inferences and Metacognitive Judgment Calibration. *Frontiers in Psychology*, 6.

Gillies, D. (2000). Varieties of propensity. *The British Journal for the Philosophy of Science*, *51*(4), 807–835.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, *2*(1), 83–115.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102, 684–704.

Girotto, V. & Gonzalez, M. (2001) Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, 78, 247–76.

Hafenbrädl, S., & Hoffrage, U. (2015). Toward an ecological analysis of Bayesian inferences: how task characteristics influence responses. *Frontiers in Psychology*, *6*.

Hoffrage, U., Krauss, S., Martignon, L., & Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Frontiers in Psychology*, *6*.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*(4), 332–340.

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, 62–88.

Johnson, E. D., & Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Frontiers in Psychology, 6.*

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.

Khan, A., Breslav, S., Glueck, M., & Hornbæk, K. (2015). Benefits of visualization in the mammography problem. *International Journal of Human-Computer Studies*, *83*, 94–113.

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.

Mandel, D. R., & Navarrete, G. (2015). Editorial: Improving Bayesian Reasoning: What Works and Why? *Frontiers in Psychology*, *6*.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive psychology*, *25*(4), 431-467.

McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological bulletin*, 143(12), 1273.

Navarrete, G., Correia, R., Sirota, M., Juanchich, M., and Huepe, D. (2015). Doctor, what does my positive test mean? From Bayesian textbook tasks to personalized risk communication. *Frontiers in Psychology* 6, 1–6.

Ottley, A., Peck, E. M., Harrison, L. T., Afergan, D., Ziemkiewicz, C., Taylor, H. A., ... & Chang, R. (2016). Improving Bayesian reasoning: the effects of phrasing, visualization, and spatial ability. *IEEE transactions on visualization and computer graphics*, *22*(1), 529–538.

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review*, *19*(3), 528–534.

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.

Pighin, S., Tentori, K., & Girotto, V. (2017). Another chance for good reasoning. *Psychonomic Bulletin & Review*, 1–8.

Reyna, V. F. (2004). How people make decisions that involve risk: A dual-processes approach. *Current directions in psychological science*, *13*(2), 60-66.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*(3), 380.

Sirota, M., Juanchich, M., & Hagmayer, Y. (2014a). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychonomic bulletin & review*, *21*(1), 198-204.

Sirota, M., Kostovičová, L., & Juanchich, M. (2014b). The effect of iconicity of visual displays on statistical reasoning: evidence in favor of the null hypothesis. *Psychonomic Bulletin & Review*, 21(4), 961–968.

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.

Appendix

Problems presented in each format in each experiment (original in Spanish).

<table>
<tr><td colspan="3">MAMMOGRAM PROBLEM (Experiment 1)</td></tr>
<tr><td>Icon Array (IA): The following figure[1] shows the prevalence of breast cancer among women over 50 who participate in routine screening, as well as the results of the mammogram for women who have and do not have breast cancer.</td><td>Natural Frequencies (NF): Among 100 women over 50 who participate in routine screening, 4 have breast cancer. 3 of the 4 women with breast cancer and 12 of the 96 women without breast cancer receive a positive mammogram.</td><td>Percentages (PE): Among the women over 50 who participate in routine screening, 4% have breast cancer. 75% of the women with breast cancer and 12% of the women without breast cancer receive a positive mammogram.</td></tr>
<tr><td colspan="3">Posterior probability question: Imagine a friend at that age receives a positive mammogram. Based on the above information, what is the probability of her having breast cancer?
(NF and IA versions prompted a “X of Y” response; PE version prompted a “%” response)</td></tr>
<tr><td colspan="3">HYPERTENSION PROBLEM (Experiment 1)</td></tr>
<tr><td>IA: The following figure[1] shows the prevalence of hypertension among women over 40 who participate in routine screening, as well as the type of diet followed by women who have and do not have hypertension.</td><td>NF: Among 100 women over 40 who participate in routine screening, 20 have hypertension. 12 of the 20 women with hypertension and 24 of the 80 women without hypertension follow a sodium-rich diet.</td><td>PE: Among the women over 40 who participate in routine screening, 20% have hypertension. 60% of the women with hypertension and 30% of the women without hypertension follow a sodium-rich diet.</td></tr>
<tr><td colspan="3">Posterior probability question: Imagine a friend at that age follows a sodium-rich diet.
Based on the above information, what is the probability of her having hypertension?
(NF and IA versions prompted a “X of Y” response; PE version prompted a “%” response)</td></tr>
<tr><td colspan="3">MAMMOGRAM PROBLEM (Experiment 2)</td></tr>
<tr><td colspan="2">IA: The following figure[1] shows the prevalence of breast cancer among women over 50 who participate in routine screening, as well as the results of the mammogram for women who have and do not have breast cancer.</td><td>NF: Among 100 women over 50 who participate in routine screening, 4 have breast cancer and 96 have not breast cancer. 3 of the women with breast cancer and 12 of the women without breast cancer receive a positive mammogram.</td></tr>
<tr><td colspan="3">Aligned question[2] (probability of the datum): Based on the above data, what is the probability of a woman at that age receiving a positive mammogram? (X of Y)
Misaligned question (posterior probability): (the same as in Experiment 1)</td></tr>
<tr><td colspan="3">HYPERTENSION PROBLEM (Experiment 2)</td></tr>
<tr><td colspan="2">IA: The following figure[1] shows the prevalence of hypertension among women over 40 who participate in routine screening, as well as the type of diet followed by women who have and do not have hypertension.</td><td>NF: Among 100 women over 40 who participate in routine screening, 10 have hypertension and 90 have not hypertension. 8 of the women with hypertension and 16 of the women without hypertension follow a sodium-rich diet.</td></tr>
<tr><td colspan="3">Aligned question[2] (probability of the datum): Based on the above data, what is the probability of a woman at that age following a rich-sodium diet? (X of Y)
Misaligned question (posterior probability): (the same as in Experiment 1)</td></tr>
</table>

[1]Corresponding icon array was presented (see an example in Figure 2) [2]Alignemnet was manipulated regarding the relational match between presented and requested relations in verbal formats (see further details in Experiment 2)

Figure captions

Figure 1. Schematic representation of the numerical relationships presented in the appendix (mammogram problem, NF format). Note that the difficulty of calculating the posterior reference class does not stem from the addition of the two focal subsets, but on the role change from subset to reference class (in bold numbers relevant for the posterior ratio; further details can be read in Experiment 2).

Figure 2. Iconic representation of the statistics presented in the appendix (mammogram problem; original version in Spanish and in color).

Figure 3. Percentage of participants who correctly solved none, one, or both problems in each group of Experiment 1.

Figure 4. Posterior probability estimates for NF and PE problems of Experiment 1.

Figure 5. Percentage of participants who correctly solved none, one, or both problems in each group (IA: icon arrays; NF: natural frequencies) and for each question in Experiment 2 (alignment refers to the relational match between the question and the text of NF problems; aligned question: probability of the datum; misaligned question: posterior probability).