What facilitates Bayesian reasoning? A crucial test of ecological rationality versus nested sets hypotheses

Gary Brase¹

Published online: 3 September 2020
© The Psychonomic Society, Inc. 2020

Abstract

Different theoretical views about Bayesian reasoning (ecological rationality and nested sets views) both claim support from results showing that natural sampling, whole numbers, and pictorial representations help with reasoning performance, although they differ in explaining how those results occur. Three studies (total $N = 653$) use minimally different numerical presentation formats—varying the singular or plural tense of the context story topic—and presence or absence of an additional icon array picture, to better understand the mechanisms driving these reasoning performance results. Plural wording, indicating a conceptual aggregation (i.e., frequencies) rather than just numerical whole numbers, consistently boosted performance. Icon arrays, in contrast, were helpful only when alongside single-tense information. These results fit more consistently with an ecological rationality view which has long argued that the mind is adapted to work best with frequentist information.

Keywords Bayesian Reasoning · Presentation Formats · Natural Frequencies · Pictorial Representation

Introduction

Certain key presentation techniques are effective for improving Bayesian reasoning (revising the likelihood of an event happening, given new information): using numbers which are naturally sampled frequencies and the use of pictorial representations of problems (McDowell & Jacobs, 2017). These techniques are important because Bayesian reasoning is prevalent across a wide range of everyday contexts, including legal, medical, and business settings (e.g., Lindsey, Hertwig, & Gigerenzer, 2003; Garcia-Retamero & Hoffrage, 2013; Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Hoffrage, Hafenbrädl, & Bouquet, 2015). Beneath these practical tips, though, is an ongoing theoretical debate about the underlying cognitive abilities foundational to that Bayesian reasoning performance.

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13423-020-01763-2>) contains supplementary material, which is available to authorized users.

✉ Gary Brase
gbrase@ksu.edu

¹ Department of Psychological Sciences, Kansas State University, Manhattan, KS, USA

The theoretical debate regarding Bayesian reasoning performance has involved two general views, referred to here as the *ecological rationality view* and the *nested sets view*. The ecological rationality view emphasizes that the human mind should be expected to be better at tasks (including Bayesian reasoning) when those tasks are presented in ways consistent with the natural and evolutionary ecology of the real-world environment. This perspective includes considerations of both the mind being evolved to expect certain types of information (Brase, Cosmides, & Tooby, 1998) and the mind being “fast and frugal”; taking advantage of structural regularities and intercorrelations which occur in the environment generally (whether evolutionarily recurrent or modern; Gigerenzer & Hoffrage, 1995; Gigerenzer, Todd, & the ABC Research Group, 1999). In contrast, the nested sets view emphasizes an ability to perceive the nested sets structure of the reasoning task, without reference to evolutionary or ecological factors having shaped the functional structure of the human mind. This view is consistent with the idea of a nested sets representation being a cognitive ability on its own (e.g., Sloman, Over, Slovak, & Stibel, 2003), incorporated within a more general mental models framework (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999), or within a more general dual systems framework (e.g., Barbey & Sloman, 2007). The distinction between ecological rationality and the nested sets views can be illustrated

and clarified by reviewing some previous research on task presentation formats, and how these findings were received by the respective views.

Numerical format and structure

Presenting numerical information as naturally sampled frequencies significantly improves Bayesian reasoning performance, as compared to the use of other formats (Gigerenzer & Hoffrage, 1995; see McDowell & Jacobs, 2017 for a meta-analysis). Naturally sampled frequencies, often shortened to just “natural frequencies,” are whole numbers which are organized into non-normalized categories and sub-categories (Kleiter, 1994). Thus, there are two features of natural frequencies: the numerical format of frequencies and the natural sampling framework within which frequencies are situated.

The ecological rationality view understands these results by recognizing that the ecology of the natural environment – at least as it is experienced by humans— largely consists of objects, events, and locations which tend to occur or not occur in countable ways. Given at least some memory of the past, the mind should thus be evolved to expect frequencies of objects, events, and locations as the quantitative input for most decisions (Brase et al., 1998). Furthermore, as that *natural sampling* information is experienced in a natural environment, it will reflect the structure of that environment.

The mind may organize this naturally sampled information into categories, and some of these categories may be subsets of larger categories (e.g., frequencies of observed people with a disease may be a subset of people seen to have a particular symptom). This organization of natural sampling information into frequencies (i.e., natural frequencies) simplifies the process of Bayesian reasoning by using non-normalized numbers that implicitly incorporate base-rate information, a fact that is agreed upon by both views (e.g., Gigerenzer & Hoffrage, 1995; Sirota, Kostovičová, & Vallée-Tourangeau, 2015). The use of externally represented frequencies (i.e., written numbers) can facilitate this process by circumventing issues such as memory demands and possible individual differences in sampling experiences, but also can be problematic (e.g., if the represented frequencies are normalized, possibly yielding a nested set configuration without a natural sampling structure).

The nested sets view, however, interprets the underlying reason for improved reasoning performance as due to a psychological process of helping people to “see” the nested set structure of the reasoning task. This has been variously referred to as promoting “clarity” or as a “partative formulation” (Macchi, 2000), as facilitating the “construction of a set inclusion mental model” (Evans, Handley, Perham, Over, & Thompson, 2000), or

as making nested relationship “opaque” or “transparent” (Sloman et al., 2003). Although no deeper insight into the computational underpinnings of this explanation has been offered, a common thread across these views is the use of some version of a natural sampling framework (Brase, 2002, 2008, 2014; Hoffrage, Gigerenzer, Krauss, & Martignon, 2002). The nested sets view concurrently rejects the ecological rationality view that the use of frequencies is a critical part of the facilitating structure, and argues that, for example, “the specific role of natural frequencies in the facilitation disappears when the comparative task features an identical partitive structure accompanied with unambiguous cues leading to similar problem representations (e.g., set operations)” (Sirota et al., 2015, p. 1472).

A key distinction to note for the present research is that the ecological rationality view considers the use of numerical frequencies as reflecting a deeper frequentist conception of the world in terms of aggregated items, whereas the nested sets view considers the use of frequencies as one of many levers to induce a nested set problem representation.

Pictorial representations

Adding a pictorial representation of task information often improves Bayesian reasoning performance (e.g., Garcia-Retamero & Hoffrage, 2013; Ichikawa, 1989; Sloman et al., 2003). More specifically, pictures that include arrays of icons are most helpful, whereas Venn circles or roulette-wheel diagrams help less (Brase, 2009, 2014; although also see Sirota, Kostovičová, & Juanchich, 2014 for slightly different findings).

The ecological rationality view on this finding is that it occurs because pictures, and especially icon arrays, visually emphasize the individuated and discrete items underlying the frequency information; consistent with the evolutionary and structural regularities of the real world. One might think the nested sets view predicts that Euler diagrams would be most helpful to visually emphasize nested sets relationships, and this appeared to be an initial stance (e.g., Sloman et al. 2003, Experiment 2; Barbey & Sloman, 2007, p. 248). Brase (2009) evaluated the hypotheses that Euler diagrams facilitate Bayesian reasoning, as they efficiently show the nested-sets structure of the task, but found that icon arrays were actually more effective in facilitating performance. The current position of the nested sets view is that all types of pictures are helpful and the degree to which they are helpful is limited to situations in which “the nested-sets structure of the problem was verbally concealed” (Sirota et al., 2014, p. 967) so that the picture takes up that role of revealing that structure rather than the text.

Hypotheses

These theoretical views have become difficult to cleanly dissociate from each other as research results, claimed as supportive by both sides, accumulate. Both views attribute the facilitative effects of natural sampling, frequencies, and pictures as primarily due to a common underlying psychological process but disagree on the nature of that process. The present research attempts to find a significant space between these views in terms of testable hypotheses. The studies employ a repeated measures design (multiple reasoning tasks), in which the numerical formats—chances or frequencies—are manipulated solely by their linguistic properties. Specifically, one format is “chances” (which are of course inherently plural) used to refer to a single entity that uses the single tense (e.g., A person, a day). The other format is “frequencies” used to refer to plural entities (e.g., people, days). In other respects, such as the use of whole numbers and framing those numbers in a natural sampling organization, the two formats are the same. Thus, this is an exceptionally strong test of if a frequentist perspective *per se* facilitates Bayesian reasoning, controlling for both computational difficulty and use of whole numbers. *If there is any performance improvement for the frequency format, relative to the chances format, it would indicate a role for the conceptual status of frequencies (i.e., a plural state of units being considered) rather than merely an effect of numerical frequencies (i.e., whole numbers).* This prediction follows directly from an ecological rationality view, but is inconsistent with a nested sets view (specifically, see Girotto & Gonzales, 2001, p. 251; 2002, pp. 357–358; Pighin, Girotto, & Tentori, 2017; Pighin, Tentori, & Girotto, 2017, p. 1995).

With regards to the effects of pictorial representations there seems to be an emerging agreement that they do in fact facilitate Bayesian reasoning, but that the result can be either a separate effect or an interaction with other format factors (e.g., numerical format). Pictures tend to have less effect at higher levels of performance, suggesting a ceiling effect to the aid in eliciting (depending on the theoretical viewpoint) either a nested sets representation or ecologically rational representation. We hypothesize that there will be some beneficial effects of including icon arrays with problems, but the pattern of those effects is somewhat unclear. The relevant hypotheses therefore are that either:

- a) pictures have a differential positive impact on the lower performing format condition; or
- b) pictures have a consistent positive effects of across all format conditions.

The former prediction is based on one of the most plausible readings of the ecological rationality view, whereas the later prediction is based on one of the most plausible readings of the nested sets view.

Experiment 1

Methods

Participants

Data were collected from 352 undergraduates at a large public university. The relatively large sample size (the target sample size was 400) was chosen to ensure sufficient power to evaluate the potentially small effects of a pictorial representation added to the same problem text as comparison conditions (Brase, 2009). For their participation, students received credit toward partial fulfillment of their introductory psychology course or extra credit in an upper-level psychology course. After eliminating 48 people who did not complete the entire study, the remaining 304 participants were used for analyses. The average age of participants was 19.96 (SD of 2.21), and 89 of the participants (29.3%) were male.

Materials and procedures

The research was implemented using an online survey platform (Qualtrics), and included the material relevant for the present research as well as a number of individual difference measures (reported in Brase, *in press*). The protocol for this and all subsequent experiments were reviewed and declared exempt by the relevant Institutional Review Board.

After reading information about the study and giving informed consent to participate, a first section of the study included a few demographic questions and a second section consisted of four Bayesian reasoning tasks, which were presented either with or without a pictorial aid and used either a single tense (chances) or plural (frequencies) numerical format (see Supplemental Materials for full text of these tasks). Thus, two tasks were presented with the numerical information in single-tense “chances” format (i.e., whole numbers in a natural sampling framework, but referring to chances of a single person’s state). One of these was just text (chances a person wears glasses, given that he is a college student), whereas the other included an icon array picture (chances of a mixed weather day, given low barometric pressure). Two other tasks were presented with the numerical information in plural frequencies (whole numbers in a natural sampling framework, but referring to a population of people and using plural tense). One of these was just text (number of university applicants who will be accepted, given passing a written test), whereas the other included an icon array picture (number of people with a disease, given a positive test reaction). These tasks were given in random order.

Table 1 The percentages of correct responses for each of the four Bayesian reasoning tasks

	No icon array	With icon array
(Single event) Chances	26.6%	43.4%
(Plural) Frequencies	57.6%	50.0%

Results

Participants varied in their performances across the different Bayesian reasoning tasks (Table 1), with the weakest performance for the single-event chances and no icon array condition (26.6%), followed by single-event chances with an icon array (43.4%). A McNemar's test of correlated proportions showed this to be a significant difference (26.6% versus 43.4%; $X^2(1) = 25.253$, $p < .001$, the odds ratio is 3.125). The two best performances were for the tasks using (plural) frequencies, although the task without an icon array (57.6%) had slightly better performance than the one with an icon array (50.0%; $X^2(1) = 4.360$, $p = .037$). These were both significant increases in performance over the respective single-tense chances presentations (43.4% versus 50.0%; $X^2(1) = 4.402$, $p = .0359$ (two-tailed), odds ratio of 1.645, and 43.4% versus 57.6%; $X^2(1) = 16.800$, $p < .001$, odds ratio of 2.387).

In order to characterize the apparent interaction, reasoning task performance was considered as dummy coded dependent variables and analyzed with a 2×2 repeated measures ANOVA. A significant effect of numerical phrasing ($F(1, 303) = 74.766$, $p < .001$, $\eta_p^2 = .198$), and a non-significant icon array presence ($F(1, 303) = 3.466$, $p = .064$, $\eta_p^2 = .011$) were qualified by a significant interaction ($F(1, 303) = 31.370$, $p < .001$, $\eta_p^2 = .094$).

Of course, all these tasks also varied in their content (the context stories and the exact numbers used), so it is possible that some of these performance differences could be due to those factors. This issue will be addressed in subsequent experiments. What cannot be used to explain the differences across these tasks is the nature and organization of the numbers in each task; all these problems used whole numbers, presented in a natural sampling framework. Figures showing the distributions of all answers (given by more than 3 participants) for each problem are provided in the Supplemental Materials.

Experiment 2

The results of Experiment 1 provide evidence in support of a facilitation in Bayesian reasoning purely due to the use of plural units being considered (i.e., conceptual frequencies) relative to a single event (also expressed as whole numbers in a natural sampling framework). Furthermore, adding icon

arrays to the singular chances formatted task had a clear beneficial impact, but the same did not occur when icon arrays were added to the plural formatted task.

It is also true, however, that the Bayesian problems in Experiment 1 each had different context stories in addition to different presentation formats. It could be that there were differences in those context stories (e.g., clarity, verbal complexity, exact numbers used) that influenced performance. There are also some differences in the complexity of the tasks: the weather prediction task and the college admission task do not include false positive rates, whereas the other tasks do. Experiments 2 and 3 address this concern.

Methods

Data were collected from 177 undergraduates at a large public university, using the same recruitment process as in Experiment 1. The sample size was reduced, compared to Experiment 1, based on the effect size strength of those results (the target sample size was 200). Each experiment was conducted during a different semester, minimizing any possible overlap in participation. After eliminating people who did not complete the entire study ($n = 21$) and finished the entire study in less than five minutes ($n = 7$), the remaining 149 participants were used for analyses. The average age of participants was 19.52 (SD of 5.71), and 64 of the participants (43.0%) were male.

The research was implemented and structured the same as in Experiment 1, with the key difference being that the context stories of the tasks were systematically reversed in both numerical presentation and icon array presence. The format for responses to the Bayesian reasoning tasks were also more tightly constrained in order to improve response quality (e.g., in terms of numerals instead of text; see Supplemental Materials for full text of these tasks).

Results

Performances across the different Bayesian tasks showed the same pattern as in Experiment 1 (Table 2). The weakest performance was for the task with chances and no icon array (14.1%), followed by frequencies with no icon array (48.3%). A McNemar's test of correlated proportions showed this to be a significant difference ($X^2(1) = 39.68$, $p < .001$, the odds ratio is 9.5). The Bayesian tasks with icon arrays led to the best

Table 2 The percentages of correct responses for each of the four Bayesian reasoning tasks

	No icon array	With icon array
(Single event) Chances	14.1%	51.0%
(Plural) Frequencies	48.3%	50.3%

performances (51.0% and 50.3%), and these were not significantly different from each other or the plain frequencies condition (e.g., 48.3% versus 51.0%: $X^2(1) = 0.188$, $p = .665$, odds ratio of 1.18).

As in Experiment 1, reasoning task performance was dummy coded and analyzed with a 2×2 repeated measures ANOVA in order to characterize any interaction effect. Significant effects of numerical phrasing ($F(1, 148) = 31.345$, $p < .001$, $\eta_p^2 = .175$) and icon array presence ($F(1, 148) = 35.767$, $p < .001$, $\eta_p^2 = .195$) were qualified by a significant interaction ($F(1, 148) = 26.618$, $p < .001$, $\eta_p^2 = .162$).

Comparing the results from Experiments 1 and 2, one effect of task context story seems to exist: the college admissions context story (which was in plural format with no icon array for Experiment 1, and singular chances with an icon array in Experiment 2) appears to facilitate performance relative to other context stories. Figures showing the distributions of all answers (given by more than 3 participants) for each problem are provided in the Supplemental Materials.

Experiment 3

Although Experiment 2 replicated the results of Experiment 1 quite well, the college admissions context story seems to have some impact of performance across formats. Experiment 3 was designed to more fully deal with any context story effects on Bayesian reasoning, by using a Latin square design to counterbalance the numerical/tense formatting and the icon array presence, relative to the context story used. Additionally, the structure of the response prompt for each task, which varied a bit across Experiments 1 and 2, were all made consistent here.

Methods

Data were collected from 239 undergraduates at a large public university, using the same recruitment process as in Experiment 1. The target sample size for this experiment was the same as for Experiment 2, adjusting for incomplete participations (the target sample size was 250). After eliminating people who did not complete the entire study ($n = 39$) and finished the entire study in less than five minutes ($n = 0$), the remaining 200 participants were used for analyses. The average age of participants was 19.32 (SD of 1.88), and 67 of the participants (33.5%) were male.

The research was implemented and structured the same as in Experiments 1 and 2, but with the order of these tasks not only random for each participant, but the use of the numerical formats and pictorial aid counterbalanced across the four context stories using a Latin square design. Thus, participants were randomly assigned to one of four possible blocks of

Table 3 The percentages of correct responses for each of the four Bayesian reasoning tasks, collapsed across task contents

	No icon array	With icon array
(Single event) Chances	23.0%	43.0%
(Plural) Frequencies	44.5%	48.5%

reasoning tasks, each systematically varying the context story, format, and icon array use. As in Experiments 1 and 2, the phrasing of the problems and the question prompts were otherwise all the same (see full text and additional details in the Supplemental Materials).

Results

Participants varied in their performances across the different Bayesian reasoning tasks (Table 3), and the weakest performance was for tasks with chances and no icon array (23%), followed by chances with an icon array (43%). A McNemar's test of correlated proportions showed this to be a significant difference ($X^2(1) = 15.55$, $p < .001$, the odds ratio is 2.90). The Bayesian tasks with plural frequencies had more consistently good and similar performances (44.5% and 48.5%). The Supplemental Materials include more finely parsed results for the Bayesian reasoning tasks, broken down by all format and story context conditions, and the distributions of all answers (given by more than 3 participants) for each story context.

As in the prior experiments, reasoning task performance was dummy coded and analyzed with a 2×2 repeated measures ANOVA in order to characterize any interaction effect. Significant effects of numerical phrasing ($F(1, 199) = 19.418$, $p < .001$, $\eta_p^2 = .098$) and icon array presence ($F(1, 199) = 15.035$, $p < .001$, $\eta_p^2 = .070$) were qualified by a significant interaction ($F(1, 199) = 6.937$, $p = .009$, $\eta_p^2 = .034$).

Discussion

Frequencies (as merely plural tense conveying an aggregation) elicit better Bayesian reasoning than singular tense chances, even if those chances employ a natural sampling (i.e., non-normalized set/subset) structure and whole numbers. The key difference thus appears to be the conceptual nature of frequentist expression, as opposed to the technical nature of frequencies as a numerical format. These results are clearly consistent with an ecological rationality approach, which has long argued that the mind is adapted to work best with frequentist information (e.g., Brase et al., 1998). These results are more difficult to explain for a nested sets view without throwing a spotlight on the nature of cognitive “transparency” as a descriptive label rather than an insightful explanation. It is

nevertheless possible, though, that this debate may continue (e.g., Pighin, Tentori, & Girotto, 2017; Brase et al., 2018).

Another clear finding here was that the addition of icon arrays helped performance on Bayesian reasoning tasks, but only in conjunction with the non-facilitative singular-tense chances wording. When the Bayesian reasoning task was given in the format of aggregate (plural) frequencies, performance was consistently higher regardless of whether there was an icon array or not. Similar results have been found before (e.g., Sirota et al., 2014; Brase & Hill, 2017) but other results (e.g., Hill & Brase, 2012) found two distinct main effects for these two format manipulations. These differing results may be due, as noted by Hill and Brase (2012), to ceiling effects in Bayesian reasoning performance within this paradigm, and further research is warranted on this issue.

From an ecological rationality view, this ceiling for icon arrays helping in Bayesian reasoning suggests that static task presentation manipulations can only achieve a certain level of mediocre performance. Better performance would be predicted by even more ecologically valid presentations of information (e.g., dynamic, interactive, or experienced information; cf. Betsch, Glauer, Renkewitz, Winkler, & Sedlmeier, 2010). From a nested sets view, this ceiling for pictures helping with reasoning raises a question about if there are limits on how much “transparency” can be built into a task presentation and how to assess that metric (i.e., Why do pictures provide additional transparency for singular tense numbers, but not to plural tense numbers?). Further research, looking not only at overall reasoning performance but correlates of performance (e.g., Brase & Hill, 2017), may help to point towards if these effects are largely independent or interdependence.

Another topic for further consideration is the influence of task context stories on performance. Hafenbrädl and Hoffrage (2015) found that there were small but measurable influences on Bayesian reasoning due to task contexts (and, interestingly, these influences were stronger when the task numerical information was presented as probabilities or percentages). In the present experiments it appears that the college admissions context story may facilitate Bayesian reasoning, whereas the medical diagnosis context story may impair reasoning performance. We are still awaiting a principled framework to understand these nascent influences.

References

- Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30(3), 241–254. <https://doi.org/10.1017/S0140525X07001653>
- Betsch, T., Glauer, M., Renkewitz, F., Winkler, I., & Sedlmeier, P. (2010). Encoding, storage and judgment of experienced frequency and duration. *Judgment and Decision Making*, 5(5), 347–364.
- Brase, G. L. (2002). Ecological and evolutionary validity: Comments on Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Cavemì's (1999) mental-model theory of extensional reasoning. *Psychological Review*, 109(4), 722–728. <https://doi.org/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. <https://doi.org/10.3758/PBR.15.2.284>
- Brase, G.L. (2009). Pictorial representations and numerical representations in Bayesian reasoning. *Applied Cognitive Psychology*, 23, 369–381. <https://doi.org/10.1002/acp.1460>
- Brase, G.L. (2014). The power of representation and interpretation: Doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *Journal of Cognitive Psychology*, 26, 81–97. <https://doi.org/10.1080/20445911.2013.861840>
- Brase, G.L. (in press). Which cognitive individual differences predict good Bayesian reasoning? Concurrent comparisons of underlying abilities. *Memory & Cognition*. <https://doi.org/10.3758/s13421-020-01087-5>
- Brase, G.L., Cosmides, L. & Tooby, J. (1998). Individuation, counting, and statistical inference: The roles of frequency and whole object representations in judgment under uncertainty. *Journal of Experimental Psychology: General*, 127, 3–21. <https://doi.org/10.1037/0096-3445.127.1.3>
- 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, 577–591. <https://doi.org/10.1037/xge0000280>
- Brase, G. L., Pighin, S., & Tentori, K. (2018). (Yet) Another chance for good reasoning? A commentary and reply on Pighin, Tentori, and Girotto (2017). *Psychonomic Bulletin & Review*, 25(3), 1220–1223. <https://doi.org/10.3758/s13423-017-1314-8>
- Evans, J. S., 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. [https://doi.org/10.1016/S0010-0277\(00\)00098-6](https://doi.org/10.1016/S0010-0277(00)00098-6)
- 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., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping Doctors and Patients Make Sense of Health Statistics. *Psychological Science in the Public Interest*, 8, 53–96. <https://doi.org/10.1111/j.1539-6053.2008.00033.x>
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102, 684–704. <https://doi.org/10.1037/0033-295X.102.4.684>
- Gigerenzer, G., Todd, P. M., & The ABC Research Group (1999). *Simple heuristics that make us smart*. New York, NY: Oxford University Press.
- 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.
- Hafenbrädl, S. & Hoffrage, U. (2015). Toward an ecological analysis of Bayesian inferences: How task characteristics influence responses. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00939>

- 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, 2343–68. <https://doi.org/10.1080/17470218.2012.687004>
- Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, 84(3), 343–352. [https://doi.org/10.1016/S0010-0277\(02\)00050-1](https://doi.org/10.1016/S0010-0277(02)00050-1)
- Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015). Natural Frequencies Facilitate Diagnostic Inferences of Managers. *Frontiers in Psychology*, 6: 642. <https://doi.org/10.3389/fpsyg.2015.00642>
- Ichikawa, S. (1989). The role of isomorphic schematic representation in the comprehension of counterintuitive Bayesian problems. *Journal of Mathematical Behavior*, 8, 269–281.
- Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J.-P. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, 106, 62–88. <https://doi.org/10.1037/0033-295X.106.1.62>
- Kleiter, G. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer, & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 375–388). New York: Springer.
- Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics*, 43, 147163. <http://www.jstor.org/stable/29762803>.
- 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>
- McDowell, M. & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, 143, 1273–1312. <https://doi.org/10.1037/bul0000126>.
- Pighin, S., Girotto, V., & Tentori, K. (2017a). 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., Tentori, K., & Girotto, V. (2017b). Another chance for good reasoning. *Psychonomic Bulletin & Review*, 24, 1995–2002. <https://doi.org/10.3758/s13423-017-1252-5>
- 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. <https://doi.org/10.3758/s13423-013-0555-4>
- Sirota, M., Kostovičová, L. & Vallée-Tourangeau, F. (2015). How to train your Bayesian: A problem-representation transfer rather than a format-representation shift explains training effects. *The Quarterly Journal of Experimental Psychology*, 68, 1–9. <https://doi.org/10.1080/17470218.2014.972420>
- Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, 91, 296–309. [https://doi.org/10.1016/S0749-5978\(03\)00021-9](https://doi.org/10.1016/S0749-5978(03)00021-9)

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.