How the level of inference in visualizations influences participants' performance in Bayesian reasoning

Michael Rößner[1], Theresa Büchter[2] and Nicole Steib[3]
[1]Paderborn University, Germany; [2]University of Kassel, Germany; [3]University of Regensburg, Germany
roessner@math.uni-paderborn.de

We conducted a study with 2,400 participants that had to solve six Bayesian reasoning tasks in one out of eight different visualization types (no visualization, regular 2×2 table, graphical 2×2 table, unit square, implicit tree diagram, explicit tree diagram, double tree, net diagram) in a probability, proportion or frequency format. The aim of the study was to investigate whether the levels of inference have an influence on the participants' performance in the tasks. The level of inference is characterized by the number of mental steps that are needed to arrive at the correct solution, which vary between the visualization types. The results show that the levels of inference indeed influence performance. This can be used to teach students to adaptively and flexibly use probabilistic visualizations for different types of tasks.

INTRODUCTION AND THEORETICAL BACKGROUND

Conditional probabilities are difficult to calculate and interpret, especially in *Bayesian reasoning tasks*, that means, in tasks with two binary features, which ask for conditional probabilities (Zhu & Gigerenzer, 2006). However, a thorough understanding of these tasks (see Table 1 for an example) can become crucial in order to make informed decisions in certain situations, for example in medicine or law (Operskalski & Barbey, 2016; Lindsey, 2003).

In the past decades of research on Bayesian reasoning, two strategies have been identified that can help to better understand and solve this type of tasks: 1) using natural frequencies instead of probabilities and 2) using a (provided) visualization (Gigerenzer & Hoffrage, 1995; McDowell & Jacobs, 2017).

Natural frequencies

In their seminal paper, Gigerenzer and Hoffrage (1995) showed that Bayesian reasoning tasks are solved correctly by a higher percentage of people, if the problem is formulated in natural frequencies (e.g., "80 out of 100 people…") instead of probabilities (e.g., "80% probability for a person to…", see also Table 1). These findings have been replicated and discussed multiple times and are usually explained by the fact that natural frequencies facilitate the calculation and make the nested-set structure of the situation more accessible (McDowell & Jacobs, 2017; Sloman, 2003).

Table 1. The famous mammography problem in the probability and frequency format (Eddy, 1982; adapted).

Probabilities	**Frequencies**
The probability of breast cancer is 2% for a woman who participates in routine screening.	200 out of 10,000 women have breast cancer.
If a woman has breast cancer, the probability is 80% that she will have a positive test result.	160 out of 200 women with breast cancer receive a positive test result.
If a woman does not have breast cancer, the probability is 10% that she will have a positive test result.	980 out of 9,800 women without breast cancer receive a positive test result.
Question: What is the probability that a woman who receives a positive test result actually has breast cancer?	**Question:** How many of the women who receive a positive test result actually have breast cancer?
Answer: ≈14%	**Answer:** 160 out of 1,140

In: Birk, L., Loth, G., Jotzo, L., Binder, K., & Frischemeier, D. (Eds.) (2026). *Statistics and Data Science Education in STEAM. Proceedings of the Satellite Conference of the International Association for Statistical Education (IASE), September/October 2025, Münster, Germany*. © ISI/IASE.

Visualizations

It has also been shown that people benefit from provided and already filled visualizations like tree diagrams, especially if the visualization is based on natural frequencies and not probabilities (Binder et al., 2015; McDowell & Jacobs, 2017). However, participants' performance and strategies in Bayesian reasoning tasks depend on the type of visualization that is used, and some visualizations do not seem to help at all (Eichler et al., 2020; Binder et al., 2015; Brase, 2009).

Levels of inference

Visualizations for situations with two binary events differ in the way they present information. Tree diagrams, for example, present a situation in a hierarchical way (i.e., conditional probabilities are only displayed in one direction), whereas 2×2 tables display no conditional, but all joint probabilities (Binder et al., 2020). Furthermore, from some visualizations (if filled completely) the answer to a Bayesian reasoning task can be directly read off (e.g., for net diagrams and double trees), while other visualizations require a calculation with multiple steps (e.g., tree diagram and unit squares). This number of arithmetic or mental steps that are needed to calculate the correct solution for given task, is expressed by the *level of inference* (Ayal & Beyth-Marom, 2014; Binder et al., 2023). A low level of inference means that the solution is directly given in the visualization, whereas a medium and high level of inference means that the number of steps to get to the correct solution are one and more than one, respectively. The level of inference does not only depend on the specific visualization, but also on the numerical format that is used (based on probabilities or natural frequencies). Table 2 gives an overview of the levels of inference for all eight visualizations that were studied in this paper and both numerical formats.

Table 2. Levels of inference for the eight visualizations used in the study. Unless otherwise noted, the examples refer to situation and the PPV question in Table 2 and the visualizations in Figures 1 and 2.

Level of inference	Low	Medium	High
Probabilities			
Visualizations	Double tree Net diagram	Regular 2×2 table Graphical 2×2 table	Text Unit square Implicit tree diagram Explicit tree diagram
Characteristics	Solution can be read off directly:	One elementary arithmetic operation is required:	More than one elementary arithmetic operation is required:
Example	32.1%	$\frac{8.5\%}{26.5\%}$	$\frac{10\% \cdot 85\%}{10\% \cdot 85\% + 20\% \cdot 90\%}$
Frequencies			
Visualizations	Regular 2×2 table Graphical 2×2 table Double tree Net diagram	Text (PPV question) Unit square Implicit tree diagram Explicit tree diagram	Text (NPV question)
Characteristics	Solution can be read off directly:	One elementary arithmetic operation is required:	More than one elementary arithmetic operation is required:
Example	85 out of 265	85 out of (85 + 180)	(900 – 180) out of ((900 – 180) + (100 – 85))

A No visualization (text)

A person is ill with a probability of 10.0%.

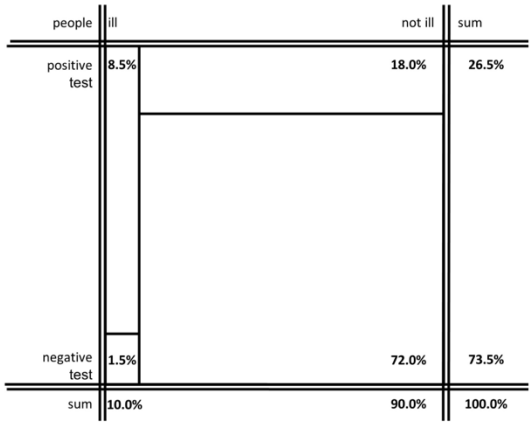
An ill person has an 85.0% probability of receiving a positive test result in the medical test.

A person who is not ill still has a 20.0% probability of testing positive in the medical test.

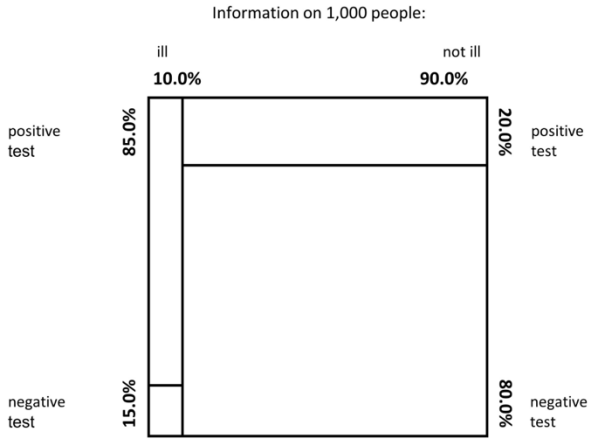
B Regular 2×2 table

people	positive test	negative test	
ill	**8.5%**	**1.5%**	**10.0%**
not ill	**18.0%**	**72.0%**	**90.0%**
	26.5%	**73.5%**	**100.0%**

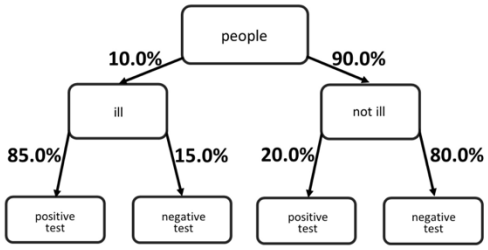
C Graphical 2×2 table



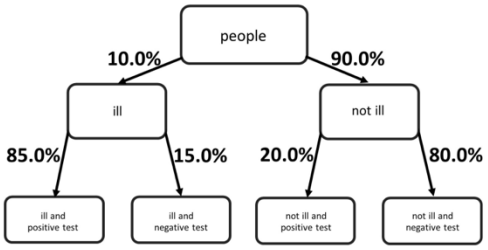
D Unit square



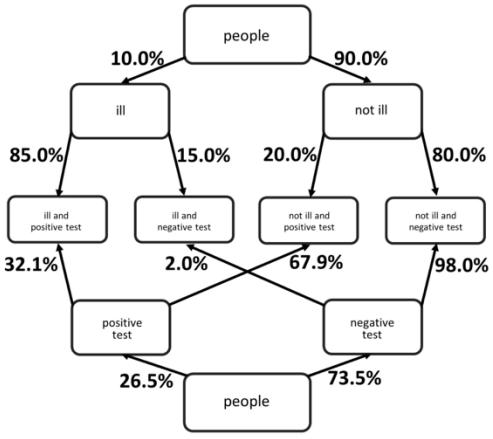
E Implicit tree diagram



F Explicit tree diagram



G Double tree



H Net diagram

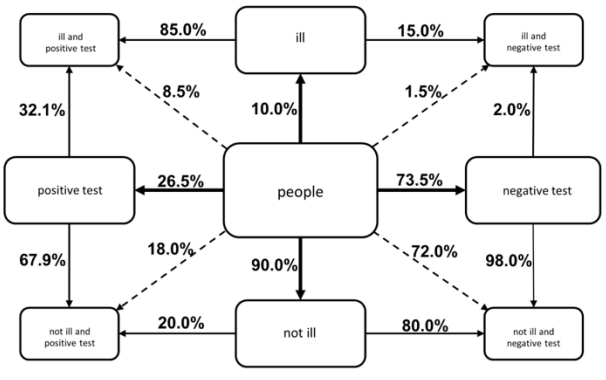


Figure 1. Visualization types (percentage format) for the task in Table 3.

A No visualization (text)

100 out of 1,000 people are ill.

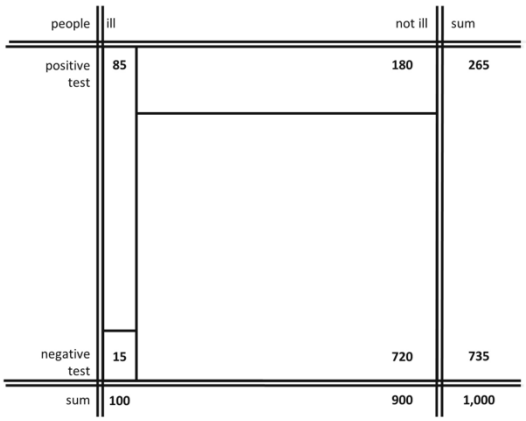
85 out of the 100 ill people get a positive test result in the medical test.

180 out of the 900 people who are not ill nevertheless test positive in the medical test.

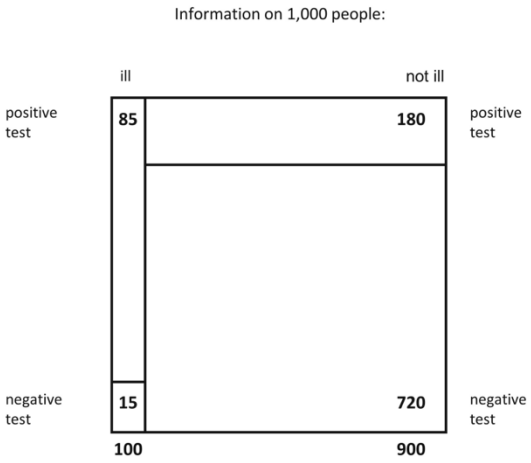
B Regular 2×2 table

people	positive test	negative test	
ill	**85**	**15**	**100**
not ill	**180**	**720**	**900**
	265	**735**	**1,000**

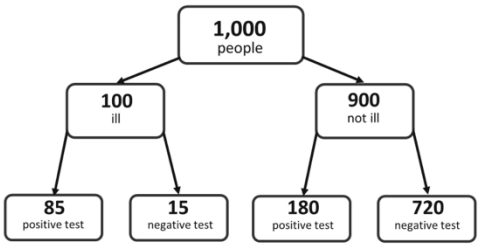
C Graphical 2×2 table



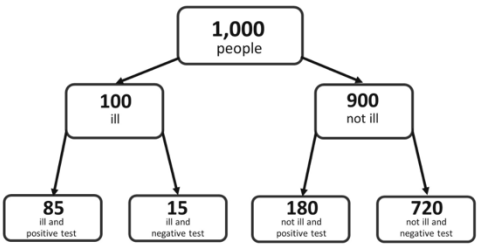
D Unit square



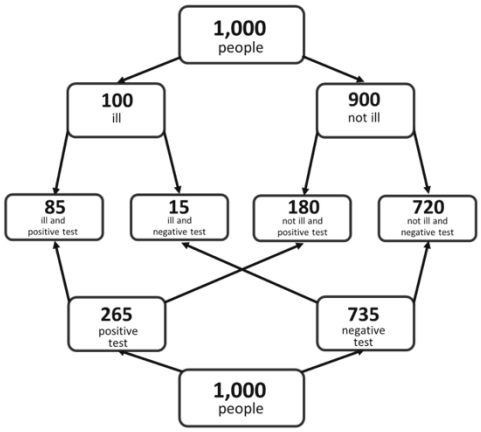
E Implicit tree diagram



F Explicit tree diagram



G Double tree



H Net diagram

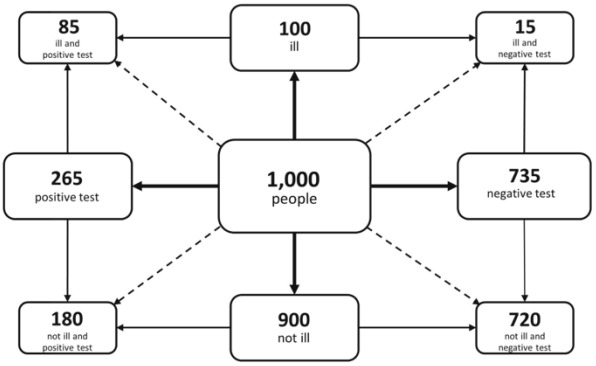


Figure 2. Visualizations (frequency format) for the task in Table 3.

RESEARCH QUESTIONS AND METHOD

In order to study whether the level of inference has an influence on participants' performance in Bayesian reasoning tasks, we conducted a study in which participants had to answer six Bayesian reasoning tasks each in a paper and pencil test in an 8×3 between subject design.

- Eight visualization types: no visualization (text only), regular 2×2 table, graphical 2×2 table, unit square, implicit tree diagram, explicit tree diagram, double tree, net diagram (see Figures 2 and 3)
- Three combinations of visualization format and question format: probabilities (probability-based visualization and probability question), natural frequencies (natural frequency-based visualization and natural frequency question), proportions (natural frequency-based visualization and proportion question)

The participants were equally distributed over the 24 conditions, leading to 100 participants per condition.

Participants

Students from several German universities and from all semesters in all fields of disciplines were recruited to participate in the study via flyers and announcements in lectures. The only requirements were that candidates had to be enrolled in a university program and be at least 18 years old. Of the 2,400 participants, 777 identified as male, 1,587 as female and 36 as non-binary. Their ages ranged from 18 to 69, with an average of 22.65 (SD = 4.18). The study was conducted in several units from April 2023 until summer 2024 with relatively small groups of about 5 to 30 people.

Materials

We used 12 different tasks, that differed by context, number set and question, because these factors have been suspected to influence the performance in Bayesian reasoning tasks (e,g., see Pighin et al., 2024; Binder et al., 2015):

- Two different contexts: medical and fairy tale
- Three different sets of numbers: varying numerical values of the given information.
- Two questions: positive predictive value (PPV) and negative predictive value (NPV)

Participants received six of these tasks on paper (one sheet per task) to enable making notes and simultaneously saw the question in a digital survey using their own laptops or tablets. The final answers were entered into the digital survey. Participants always had to hand in the previous task (on paper) before continuing to the next one. Table 2 shows one example of a Bayesian reasoning task, that was used in the study. All eight visualizations are depicted in Figures 1 and 2.

RESULTS

Figure 3 shows the participants' performance in the Bayesian reasoning tasks, separated by the combination of visualization and question format, and level of inference. For all formats, the proportion of correctly solved tasks is highest, if a visualization with a low level of inference was provided and lowest, if a visualization with a high level of inference was provided. We used generalized linear mixed models to predict the probability for a correct solution for each format and found that the differences between the levels of inference were statistically significant in all formats.

However, one more thing is striking: in the probability format, even with visualizations with a low level of inference (i.e., the correct solution can be read off directly!) only 37.8% of the participants were able to identify the correct solution.

CONCLUSION

Performance in Bayesian reasoning tasks depends 1) on the level of inference and 2) on the formats of both the question and the visualization. These findings imply that students should be supported by learning how to translate the given numerical format of the question (e.g., from probabilities to natural frequencies; see Feufel et al., 2023) and/or to reduce the level of inference (e.g., by sketching visualizations with a low level of inference to solve such problems; see Steib et al., 2025). Still, as we focused on tasks with provided visualizations, future research might be necessary to relate

these results to other findings that focus more on the construction process of visualizations (Rößner et al., 2025).

Table 3. Exemplary task (medical context, originally German, translated to English). Note that for each task only one of the two questions (PPV or NPV) was used.

Visualization format	**Percentages**	**Frequencies**	**Frequencies**
Question format	**Probabilities**	**Frequencies**	**Proportions**
Introduction	The following information is known about a medical test:		
Given information (presented in text as in this table or in one of the seven visualizations displayed in figures 1 and 2)	A person is ill with a probability of 10.0%. An ill person has an 85.0% probability of receiving a positive test result in the medical test. A person who is not ill still has a 20.0% probability of testing positive in the medical test.	100 out of 1,000 people are ill. 85 out of the 100 ill people get a positive test result in the medical test. 180 out of the 900 people who are not ill nevertheless test positive in the medical test.	100 out of 1,000 people are ill. 85 out of the 100 ill people get a positive test result in the medical test. 180 out of the 900 people who are not ill nevertheless test positive in the medical test.
Question (PPV)	What is the probability that a person with a positive test result is actually ill?	How many people with a positive test result are actually ill?	What proportion of people with a positive test result are actually ill?
Question (NPV)	What is the probability that a person with a negative test result is actually not ill?	How many people with a negative test result are actually not ill?	What proportion of people with a negative test result are actually not ill?
Answer template	(percentage, rounded to one decimal place): ___.__%	_____ out of _____	(as a fraction): ▭/▭

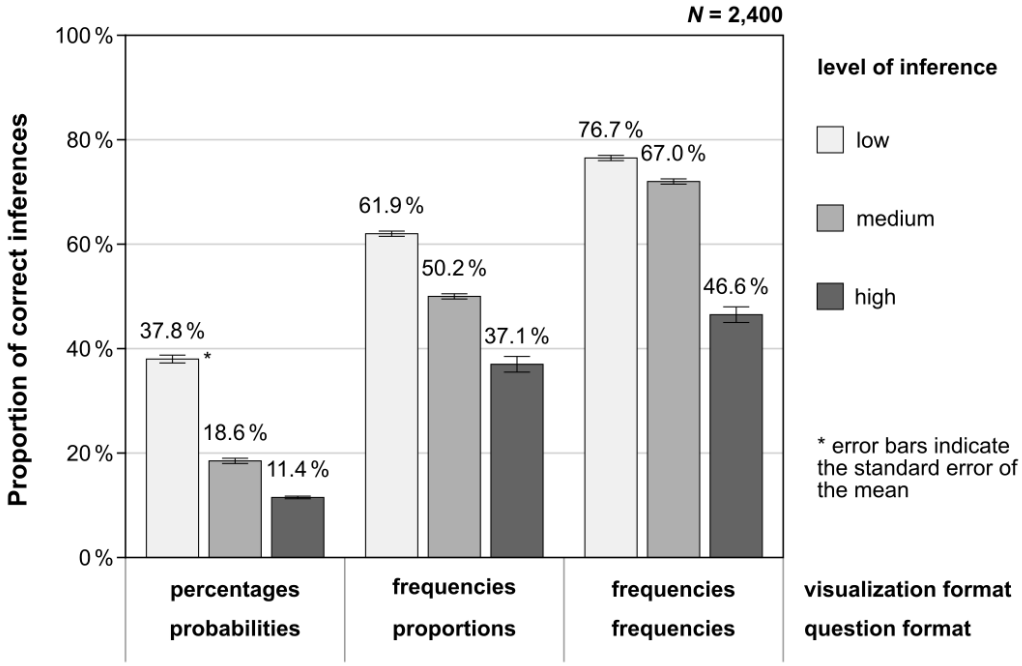


Figure 3. Proportion of correct inferences, separated by the combination of visualization and question format and the level of inference.

REFERENCES

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

Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information – an empirical study on tree diagrams and 2 × 2 tables. *Frontiers in Psychology, 6*, Article 1186. https://doi.org/10.3389/fpsyg.2015.01186

Binder, K., Krauss, S., & Wiesner, P. (2020). A new visualization for probabilistic situations containing two binary events: The frequency net. *Frontiers in Psychology, 11*, Article 750. https://doi.org/10.3389/fpsyg.2020.00750

Binder, K., Steib, N., & Krauss, S. (2023). Von Baumdiagrammen über Doppelbäume zu Häufigkeitsnetzen – kognitive Überlastung oder didaktische Unterstützung? *Journal für Mathematik-Didaktik, 44*, 471–503. https://doi.org/10.1007/s13138-022-00215-9

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology, 23*(3), 369–381. https://doi.org/10.1002/acp.1460

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 249–267). Cambridge University Press.

Eichler, A., Böcherer-Linder, K., & Vogel, M. (2020). Different visualizations cause different strategies when dealing with Bayesian situations. *Frontiers in Psychology*, *11*, Article 1897. https://doi.org/10.3389/fpsyg.2020.01897

Feufel, M. A., Keller, N., Kendel, F., & Spies, C. D. (2023). Boosting for insight and/or boosting for agency? How to maximize accurate test interpretation with natural frequencies. *BMC Medical Education, 23*(1), Article 75. https://doi.org/10.1186/s12909-023-04025-6

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

Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics, 43*(2), 147–163. https://www.jstor.org/stable/29762803

McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin, 143*(12), 1273–1312. https://psycnet.apa.org/doi/10.1037/bul0000126

Operskalski, J. T., & Barbey, A. K. (2016). Risk Literacy in Medical Decision-Making. *Science, 352*(6284), 413–414. https://doi.org/10.1126/science.aaf7966

Pighin, S., Filimon, F., & Tentori, K. (2024). The impact of problem domain on Bayesian inferences: A systematic investigation. *Memory and Cognition, 52*, 735–751. https://doi.org/10.3758/s13421-023-01497-1

Rößner, M., Binder, K., Geier, C., & Krauss, S. (2025). Students' performance and typical errors in filling empty probabilistic visualizations with probabilities or frequencies. *Educational Studies in Mathematics*, *120*, 137–167. https://doi.org/10.1007/s10649-024-10372-y

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, *91*(2), 296–309. https://doi.org/10.1016/S0749-5978(03)00021-9

Steib, N., Büchter, T., Eichler, A., Binder, K., Krauss, S., Böcherer-Linder, K., Vogel, M., & Hilbert, S. (2025). How to teach Bayesian reasoning: An empirical study comparing four different probability training courses. *Learning and Instruction*, *95*, Article 102032. https://doi.org/10.1016/j.learninstruc.2024.102032

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: the role of representation in mental computation. *Cognition, 98*(3), 287–308. https://doi.org/10.1016/j.cognition.2004.12.003