T(H)REE STEPS TO IMPROVE BAYESIAN REASONING

Karin Binder[1], Stefan Krauss[2], Georg Bruckmaier[1] and Jörg Marienhagen[3]
[1]Mathematics Education, University of Regensburg, Germany
[2]University of Applied Sciences Northwestern, Switzerland
[3]University Hospital Regensburg, Germany
Karin.Binder@ur.de

Physicians must frequently combine statistical information on prevalence of diseases and on medical tests according to Bayes' theorem in order to arrive at a correct diagnosis. Such decision-making processes are often associated with significant errors of judgment, which have been documented repeatedly with respect to the Bayesian "standard" task (i.e., one disease must be diagnosed based on one medical test). In the present contribution, we generalize the Bayesian reasoning paradigm to medical situations where more than just one medical test is involved and suggest three steps towards a better understanding of that generalized situation, namely, 1) replace probabilities with natural frequencies; 2) present a tree diagram containing natural frequencies; and 3) highlight the two branches of the tree that are relevant for the requested inference.

BACKGROUND

In medicine, empirical studies repeatedly show that physicians frequently struggle in their decision-making processes when statistical information on the prevalence of a disease must be combined with the sensitivities and false-alarm rates of medical tests in order to come up with what is known as the positive predictive value (Gigerenzer & Hoffrage, 1995; Hoffrage, Krauss, Martignon, & Gigerenzer, 2015). Imagine a woman who participates in a routine screening for breast cancer by having a mammogram, with the following statistical information available:

Breast cancer screening – 1-test case (probability format):

The probability of having breast cancer for a woman of a particular age group is 1%. The probability that a woman with breast cancer will have a positive mammogram is 80%. The probability that a woman without breast cancer will have a false-positive mammogram is 9.6%. What is the probability that a woman with a positive mammogram actually has breast cancer?

Most physicians assume this probability to be between 70 and 80 % (Eddy, 1982) although the correct solution is about 7.8%. Fortunately, however, there are two strategies for overcoming the occurring cognitive illusions (e.g., base rate neglect) and fostering insight into these *Bayesian "standard" tasks*, namely 1) provide statistical information with natural frequencies instead of probabilities, and 2) provide visualizations of the statistical information. Let us focus first on the strategy of using "natural frequencies". After translating probabilities into natural frequencies, the situation reads:

Breast cancer screening – 1-test case (natural frequency format):

100 out of 10,000 women of a particular age group who participate in a routine screening have breast cancer. 80 out of 100 women who participate in routine screening and have breast cancer will have a positive mammogram. 950 out of 9,900 women who participate in routine screening and have no breast cancer will have a false-positive mammogram. How many of the women who have participated in a routine screening and received positive mammograms have breast cancer?

A huge number of studies and a recent meta-analysis (McDowell & Jacobs, 2017) confirm that more people are able to find the correct solution when all statistical information is presented with natural frequencies (Gigerenzer & Hoffrage, 1995). There are multiple studies about the 1-test case, but often medical reality is less straightforward–sometimes further test results have to be integrated. Fortunately, the natural frequency format is also helpful in scenarios where more than one medical test has to be taken into account. Imagine, for example, a breast cancer screening where both a mammogram and then, as a next step, a sonogram are conducted:

Breast cancer screening – 2-test case (probability format):

The probability of breast cancer for a woman of a particular age group is 1%. The probability that a woman with breast cancer will have a positive mammogram is 80%. The probability that a woman with breast cancer will have a positive sonogram is 95%. The probability

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

that a woman without breast cancer will have a false-positive mammogram is 9.6%. The probability that a woman without breast cancer will have a false-positive sonogram is 7.8%. What is the probability that a woman with a positive mammogram and a positive sonogram actually has breast cancer?

The correct solution of this “2-test case” is 50.7%. However, when statistical information is provided in probabilities, most people again fail in their decision-making processes (Krauss, Martignon, & Hoffrage, 1999; Hoffrage et al., 2015). But here it is again possible to replace percentages with natural frequencies:

Breast cancer screening – 2-test case (natural frequency format):

100 out of 10,000 women of a particular age group have breast cancer. 80 out of 100 women with breast cancer have a positive mammogram. 76 out of 80 women with breast cancer and a positive mammogram have a positive sonogram. 950 out of 9,900 women without breast cancer have a false-positive mammogram. 74 out of 950 women without breast cancer but with a positive mammogram have a false-positive sonogram. How many of the women with a positive mammogram and a positive sonogram actually have breast cancer?

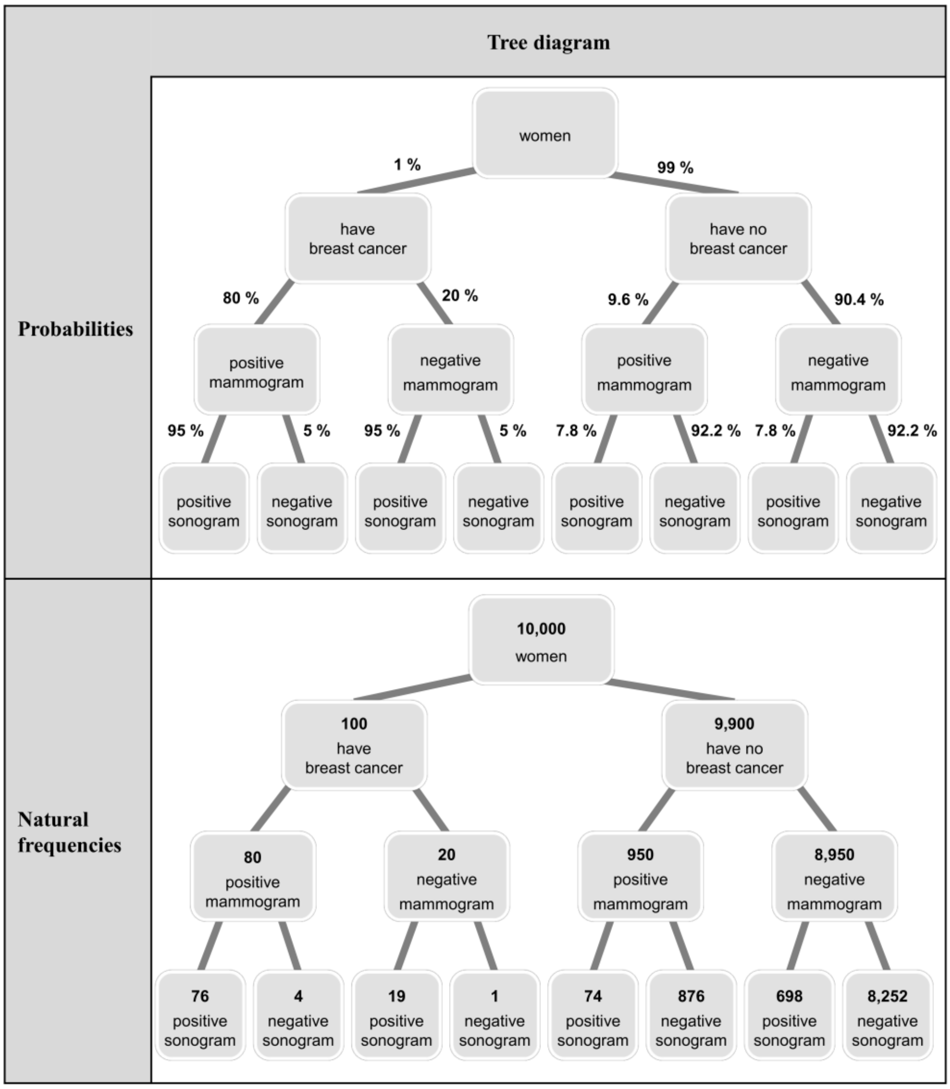

Figure 1. Natural frequency tree and probability tree for the breast cancer screening problem implemented in Study 1.

In natural frequency format, many more people are able to see that 76 out of (76+74) women actually have breast cancer (Krauss, Martignon, & Hoffrage, 1999; Hoffrage et al., 2015).

Another strategy that can improve Bayesian reasoning in the 1-test case is using visualizations, for example, tree diagrams (e.g., Binder, Krauss, & Bruckmaier, 2015; Yamagishi, 2003; see also Figure 1), 2×2 tables (e.g., Binder, Krauss, & Bruckmaier, 2015; Steckelberg et al., 2004), frequency grids (e.g., Sedlmeier & Gigerenzer, 2001; Garcia-Retamero & Hoffrage, 2013), unit squares (e.g., Böcherer-Linder & Eichler, 2017; Pfannkuch & Budgett, 2016), Euler diagrams (Sirota et al., 2014), or icon arrays (e.g., Brase, 2014; Zikmund-Fisher et al., 2014). However, it is an open question as to whether tree diagrams (or other visualizations) can also improve Bayesian reasoning in the 2-test case. In Study 1 we want to test whether trees are more helpful *instead of* or *in addition to* textual wordings, and in Study 2 we want to deepen insight into which kind of tree diagram is most helpful (see Binder, Krauss, Bruckmaier, & Marienhagen, 2018).

STUDY 1

In a paper-and-pencil questionnaire, 190 medical students from the University of Regensburg worked on two different Bayesian reasoning problems (one of them was the breast cancer screening problem with two medical tests as described above). We implemented a 3×2×2 design with the factors *presentation of information* (text only vs. tree only vs. text and tree), *information format* (probabilities vs. natural frequencies) and *context* (breast cancer screening problem vs. HIV testing problem). The implemented probability tree and natural frequency tree are depicted in Figure 1.

Each participant received one of the two problem contexts in probability format and the other in natural frequency format, with the order of context and information format varied systematically. When one of the problems the participant worked on had a certain presentation of information (e.g., text only), the other problem contained one of the other remaining types of information presentation.

Figure 2 illustrates the results of Study 1. Providing the information in natural frequency format instead of in probability format improved participants' performance substantially. Furthermore, the presentation of a tree diagram with natural frequencies was also helpful for our participants. Interestingly, it makes no difference whether the tree diagram is presented *in addition to* or *instead of* the textual statistical information. The text-and-tree version reached the highest performance rate of 48%. With the probability version, however, it does not really matter if a tree diagram is provided in addition to or instead of a textual wording. In all probability versions, participants' performance was lower than 10%.

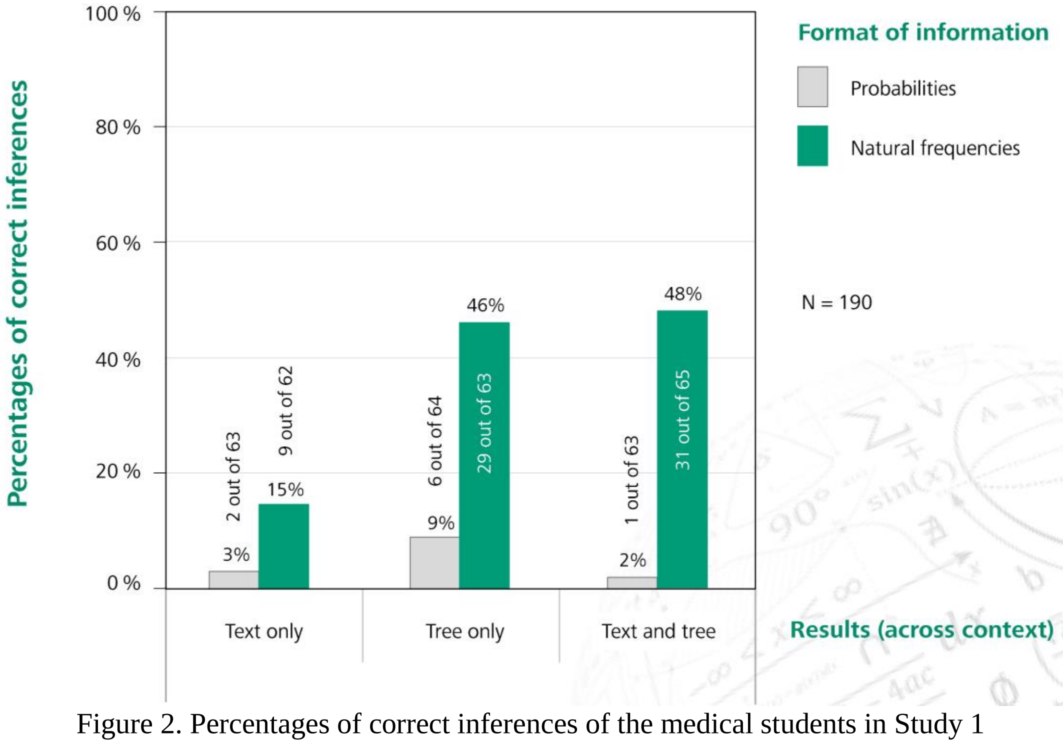

Figure 2. Percentages of correct inferences of the medical students in Study 1

Although natural frequencies and tree diagrams could improve participant performance, however, there is still room for improvement (only 48% of the participants could solve the tasks with natural frequency trees). Therefore, we analyzed two modified tree diagrams in the second study in order to improve participant performance even more.

STUDY 2

In a paper-and-pencil questionnaire, 198 medical students (none of whom had participated in Study 1) from the University of Regensburg worked on two different Bayesian reasoning problems. As in Study 1, we implemented a 3×2×2 design with the factors *presentation of information* (complete tree vs. highlighted tree vs. pruned tree), *information format* (probabilities vs. natural frequencies) and *context* (breast cancer screening problem vs. HIV testing problem). Thus this time, all versions had tree diagrams. The three different implemented natural frequency tree diagrams can be seen in Figure 3 (respective probability trees were presented accordingly).

Again, each participant received one of the two problem contexts in probability format and the other in natural frequency format, with the order of context and information format once again varied systematically. When one of the problems that a participant worked on had a certain presentation of information (e.g., complete tree diagram), the other problem contained one of the other remaining types of information presentation.

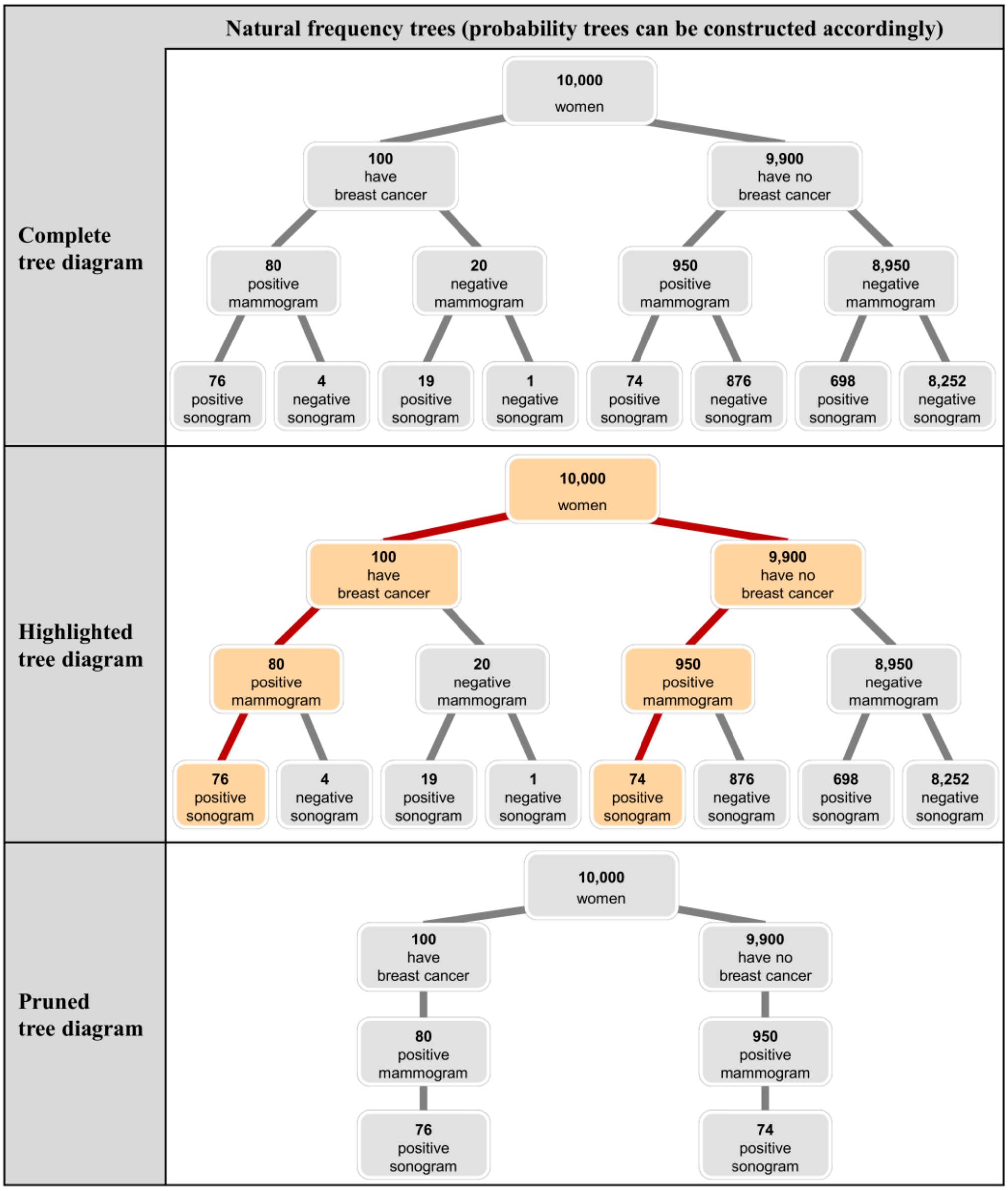

Figure 3. Natural frequency trees for the breast cancer screening problem presented in Study 2.

Figure 4 illustrates the results of Study 2. Again, natural frequency trees helped participants in their decision-making processes, whereas probability trees did not and participants' performance was very low in all probability conditions. Highlighting the two relevant branches in the natural frequency trees yields the highest performance rates. Interestingly, pruning the tree and only showing the question-related branches does not improve the performance over and above the level that was reached with a complete tree diagram. In sum, the natural frequency tree with highlighted branches was the most help to medical students.

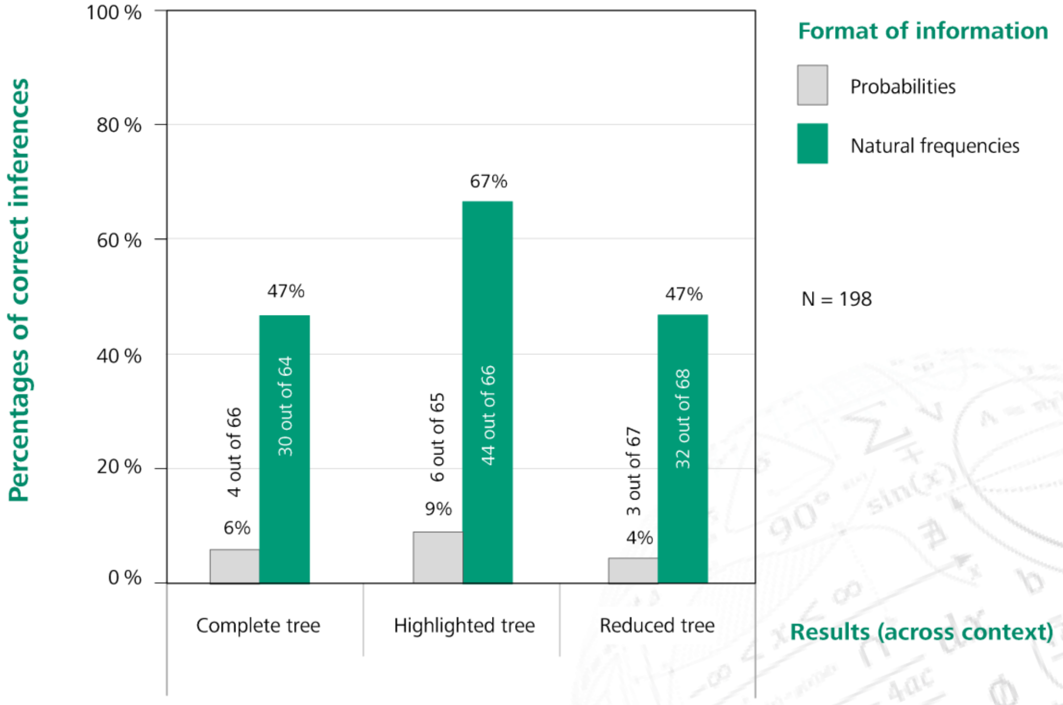

Figure 4. Percentages of correct inferences by medical students in Study 2

CONCLUSION

With respect to medical decision making, understanding the meaning of medical test results is crucial for medical students and physicians as well as for patients because it can reduce the possible harms of overdiagnosis and overtreatment (Wegwarth & Gigerenzer, 2013) but can also reduce the danger of serious diseases being overlooked.

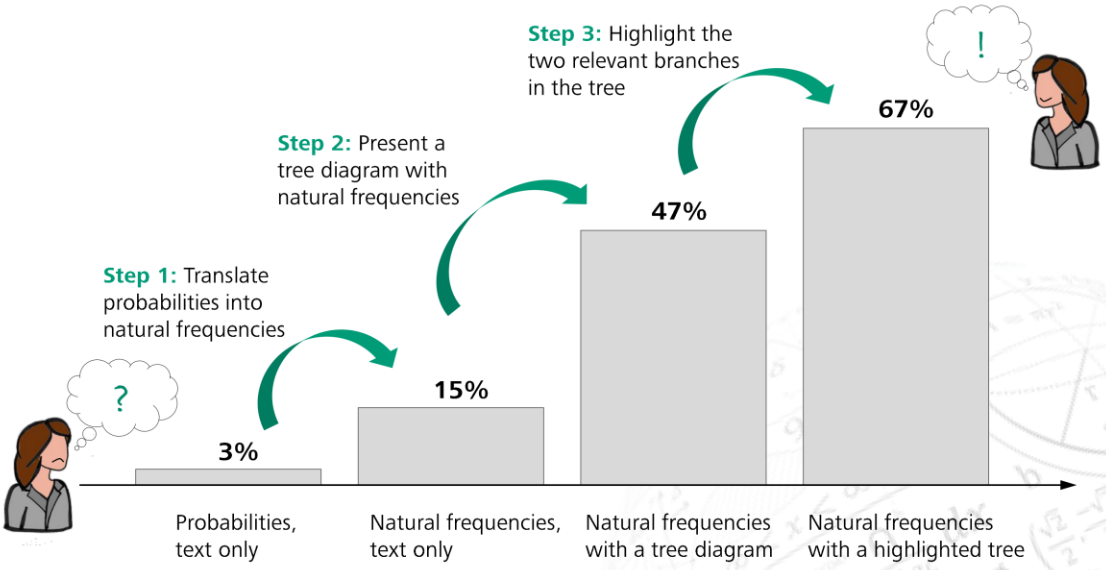

Figure 5. T(h)ree steps to improve Bayesian reasoning

It has to be noted that we did not run a training study. However, without training, 67% of the participants were able to make correct inferences in Bayesian 2-test cases using highlighted natural frequency trees. Consequently, (highlighted) natural frequency trees can be used without any prior training and therefore should be implemented in medical textbooks and statistics education materials for prospective physicians.

Figure 5 summarizes the results of Study 1 and Study 2. These results indicate that three recommendations can be made for teaching Bayesian reasoning problems: 1) Translate (conditional) probabilities into natural frequencies. 2) Present a tree diagram with natural frequencies (in addition to or instead of the textual version of the problem). 3) Highlight the two branches of the tree diagram that are relevant for answering the question.

REFERENCES

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*(1186).

Binder, K., Krauss, S., Bruckmaier, G., & Marienhagen, J. (2018). Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making. *PLoS ONE*.

Böcherer-Linder, K., & Eichler, A. (2017). The impact of visualizing nested sets. An empirical study on tree diagrams and unit squares. *Frontiers in psychology*, *7*, 2026.

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*(1), 81-97.

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.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, *102*(4), 684–704.

Hoffrage, U., Krauss, S., Martignon, L., & Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Frontiers in psychology*, *6*(1473).

Krauss, S., Martignon, L., & Hoffrage, U. (1999). Simplifying Bayesian Inference: The General Case. In N. e. a. Magnani (Ed.), *Model-based Reasoning in Scientific Discovery* (pp. 165–179), Kluwer Academic/Plenum Publishers, New York.

McDowell, M., & Jacobs, P. (2017). Meta-Analysis of the Effect of Natural Frequencies on Bayesian Reasoning. *Psychological bulletin*, *143*(12), 1273-1312.

Pfannkuch, M., & Budgett, S. (2017). Reasoning from an eikosogram: An exploratory study. *International Journal of Research in Undergraduate Mathematics Education*, *3*(2), 283-310.

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

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*(4), 961-968.

Steckelberg, A., Balgenorth, A., Berger, J., & Mühlhauser, I. (2004). Explaining computation of predictive values: 2 × 2 table versus frequency tree. A randomized controlled trial [ISRCTN74278823]. *BMC medical education*, *4*(1), 13.

Wegwarth, O., & Gigerenzer, G. (2013). Overdiagnosis and overtreatment: evaluation of what physicians tell their patients about screening harms. *JAMA internal medicine*, *173*(22), 2086-2088.

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets? *Experimental Psychology*, *50*(2), 97.

Zikmund-Fisher, B. J., Witteman, H. O., Dickson, M., Fuhrel-Forbis, A., Kahn, V. C., Exe, N. L., ... & Fagerlin, A. (2014). Blocks, ovals, or people? Icon type affects risk perceptions and recall of pictographs. *Medical decision making*, *34*(4), 443-453.