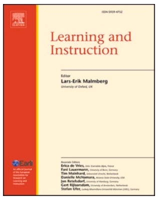
Learning and Instruction 95 (2025) 102032

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc



How to teach Bayesian reasoning: An empirical study comparing four different probability training courses

Nicole Steib[a,*], Theresa Büchter[b], Andreas Eichler[b], Karin Binder[c], Stefan Krauss[a], Katharina Böcherer-Linder[d], Markus Vogel[e], Sven Hilbert[f]

[a] Faculty of Mathematics, University of Regensburg, Universitätsstraße 31, D-93053, Regensburg, Germany
[b] Institute of Mathematics, University Kassel, Heinrich-Plett-Str. 40, D-34132, Kassel, Germany
[c] Mathematical Institute, Ludwig Maximilian University, Theresienstr. 39, D-80333, Munich, Germany
[d] Department of Mathematics Education, University of Freiburg, Freiburg, Germany, Ernst-Zermelo-Straße 1, D-79104, Freiburg, Germany
[e] Institute of Mathematics, University of Education Heidelberg, Keplerstraße 87, D-69120, Heidelberg, Germany
[f] Faculty of Psychology, University of Regensburg, Sedanstraße 1, D-93055, Regensburg, Germany

ARTICLE INFO

Keywords:
Bayesian reasoning
Training study
Double tree
Unit square
Natural frequencies
Probability tree

ABSTRACT

Background: Bayesian reasoning is understood as the updating of hypotheses based on new evidence (e.g., the likelihood of an infection based on medical test results). As experts and students alike often struggle with Bayesian reasoning, previous research has emphasised the importance of identifying supportive strategies for instruction.
Aims: This study examines the learning of Bayesian reasoning by comparing five experimental conditions: two “level-2” training courses (double tree and unit square, each based on natural frequencies), two “level-1” training courses (natural frequencies only and a school-specific visualisation “probability tree”), and a “level-0” control group (no training course). Ultimately, the aim is to enable experts to make the right decision in high-stake situations.
Sample: $N = 515$ students (in law or medicine)
Method: In a pre-post-follow-up training study, participants’ judgments regarding Bayesian reasoning were investigated in five experimental conditions. Furthermore, prior mathematical achievement was used for predicting Bayesian reasoning skills with a linear mixed model.
Results: All training courses increase Bayesian reasoning, yet learning with the double tree shows best results. Interactions with prior mathematical achievement generally imply that students with higher prior mathematical achievement learn more, yet with notable differences: instruction with the unit square is better suited for high achievers than for low achievers, while the double tree training course is the only one equally suited to all levels of prior mathematical achievement.
Conclusion: The best learning of Bayesian reasoning occurs with strategies not yet commonly used in school.

1. Introduction

“Most important decisions men make are governed by beliefs concerning the likelihood of unique events” (Tversky & Kahneman, 1973, p. 231). Today, this may be even more the case than ever, since nowadays we live in an increasingly data-driven society (Radermacher, 2022) where statistical and probabilistic information do not only govern political debates and business decisions but are also a highly relevant source for individual decision making (Galesic & Garcia-Retamero, 2010; Reyna & Brainerd, 2007). Therefore, understanding, critically evaluating and deciding on probabilistic information, are more and more understood as decisive tools for every “citizen’s need to cope with uncertainty and risk in the modern world” (Batanero & Álvarez-Arroyo, 2024, p. 5) and are also essential concepts in statistical literacy and statistics education (Burril, 2020; Burrill & Pfannkuch, 2024; Gal & Geiger, 2022).

* Corresponding author. University of Regensburg, Universitätsstraße 31, Germany.
E-mail addresses: nicole.steib@ur.de (N. Steib), tbuechter@mathematik.uni-kassel.de (T. Büchter), eichler@mathematik.uni-kassel.de (A. Eichler), karin.binder@lmu.de (K. Binder), stefan.krauss@mathematik.uni-regensburg.de (S. Krauss), boecherer-linder@math.uni-freiburg.de (K. Böcherer-Linder), vogel@ph-heidelberg.de (M. Vogel), sven.hilbert@ur.de (S. Hilbert).

https://doi.org/10.1016/j.learninstruc.2024.102032
Received 20 December 2023; Received in revised form 4 September 2024; Accepted 5 October 2024
Available online 1 November 2024

Coping with uncertainty and risk often entails handling conditional probabilities. Especially Bayesian reasoning, which is the process of updating probabilities for hypotheses based on new information to arrive at conclusions and make decisions (Reani et al., 2018), is crucial for different professions, for example, law (Lindsey et al., 2003), medicine (Gigerenzer et al., 2007) or economics (Hoffrage, Hafenbrädl, & Bouquet, 2015). For instance, a judge should be able to revise the probability of a potential guilt of an accused person based on a particular piece of evidence (e.g., a DNA-test result) and a physician has to update the probability of a certain illness in the light of medical test results (Fig. 1). Erroneous Bayesian reasoning is often reported (Binder et al., 2020; Eichler et al., 2020) even among experts (Hoffrage & Gigerenzer, 1998), and can have dramatic consequences. For example, physicians and HIV consultants often misinterpret positive test results, assuming that being infected with HIV is absolutely certain if the subject tested positive in an HIV test even though the probability in the described scenario is only about 70% (Ellis & Brase, 2015; Prinz et al., 2015). Faulty Bayesian reasoning can have tragic consequences such as overtreatment in medicine (Wegwarth & Gigerenzer, 2013) or even suicides based on wrong interpretations of test results (Stine, 1996).

Justifiably, Bayesian reasoning is also part of statistics education in universities and schools in the framework of teaching conditional probabilities (Borovcnik, 2016; Veaux et al., 2012). Although probability and statistics are increasingly included in curricula worldwide, the actual implementation in textbooks and teaching is not always satisfactory (Batanero et al., 2016; Binder et al., 2015). Therefore, research on how to teach Bayesian reasoning and avoid typical errors is relevant both for teaching probabilities at school and for training stakeholders as well.

In the present paper, we empirically examine the effects of four computer-based training courses for learning Bayesian reasoning. Participants ($N = 515$) are university students of medicine and law, thus the focus is on preparing future experts with the skills for correct Bayesian reasoning. In doing so, we additionally explore interactions with prior mathematical achievement for determining which training course is appropriate for which learner. Although our study does not directly deal with secondary teaching, our training courses could also be used in school teaching of probability.

2. Bayesian reasoning

The simplest case of Bayesian situations consists of a binary hypothesis H (e.g., ill vs. healthy) and binary information I (e.g., test positive vs. test negative; Zhu & Gigerenzer, 2006). In such situations, Bayesian reasoning implies estimating conditional probabilities (McDowell & Jacobs, 2017) based on three probabilities (i.e., base rate P(H), true-positive rate $P(I|H)$ and false-positive rate $P(I|\overline{H})$)[1] and a typical task is to assess the positive (or negative) predictive value (PPV or NPV; for an example, see Fig. 1). Mathematically, the inference can be modelled using Bayes theorem:

$$P(H|I) = \frac{P(H)\cdot P(I|H)}{P(H)\cdot P(I|H) + P(\overline{H})\cdot P(I|\overline{H})} \text{ (positive predictive value)} \qquad \text{(equation 1)}$$

Fig. 1 displays a Bayesian situation about prenatal screenings with a triple test. The triple test holds promising characteristics: the probability for a woman carrying an unborn child *with* Down syndrome to correctly test positive is 75% (true-positive rate), while the probability for a woman carrying an unborn child *without* Down syndrome to falsely test positive is only 5% (false-positive rate). Nevertheless, in this situation with a base rate of 3%, the PPV is only about 32% for a woman to *actually* carry an unborn child with Down syndrome if this woman tests positive.

Such percentages for the PPV have often been documented as unintuitively low (Hoffrage & Gigerenzer, 1998). An obstacle for correct Bayesian reasoning is the tendency to overlook the influence of the base rate for estimations of the PPV. This bias became known as *base rate neglect* in the research of Kahneman and Tversky (1982). The meta-analysis for Bayesian reasoning by McDowell and Jacobs (2017) shows that the performance for correctly assessing the PPV (or NPV) without previous instruction is only about 5%, if the statistical information in such situations is given in probabilities (Fig. 1, left). Even among experts, the performance is similarly poor, as shown in the field of law or medicine (e.g., Kurzenhäuser & Hoffrage, 2009; Lindsey et al., 2003). As a consequence, patients such as the pregnant women in the example on prenatal screenings may be misled and therefore assume a much higher degree of certainty of the positive test result than would be appropriate. Fatal consequences may be abortions and unnecessary worrying of the pregnant women (Roberts et al., 2002; West & Brase, 2023). The topic of Bayesian reasoning including such tragic consequences have even been repeatedly published in journals such as *Science* (e.g., Spiegelhalter et al., 2011; Tversky & Kahneman, 1974) or *Nature* (e.g., Goodie & Fantino, 1996), highlighting the urgent need to improve people's Bayesian reasoning abilities.

In reality, the starting point in domains such as in medicine and law is usually a representation of such a situation in probabilities with no additional visualisation of the statistical information provided (see Fig. 1, left). In principle, previous research established that fostering the understanding of Bayesian reasoning can be approached by a) using supportive representations of the Bayesian situation (see 2.1) or b) explicit training courses (see 2.2).

2.1. *Supportive representations of statistical information*

2.1.1. *Natural frequencies*

Research on decision making under uncertainty and probability teaching within the last three decades has identified several supportive representations for Bayesian reasoning.

One helpful strategy goes back to the seminal work by Gigerenzer and Hoffrage (1995) in which they introduced the so-called natural frequencies (Fig. 1, right). These can be understood as a pair of natural numbers "a out of b" with a ≤ b (Krauss et al., 2020) and relate the probabilistic information to a concrete sample of individuals (e.g., 10, 000 pregnant women) through the principle of natural sampling (Kleiter, 1994). One advantage of natural frequencies is that they simplify the calculation of the correct solution (Gigerenzer & Hoffrage, 1995; McDowell & Jacobs, 2017). Furthermore, the influence of the base rate is more palpable than with probabilities. For instance, only in the natural frequency format does it become clear why so many false-positives appear: despite the low false-positive rate of 5%, the absolute number of false positives (485) is more than double compared to the amount of true positives (225), simply because there are so many more women carrying an unborn child *without* Down syndrome (9,700, where false positives might appear) than women carrying an unborn child *with* Down syndrome (300, where true positives might appear). According to the meta-analysis, the performance in Bayesian reasoning tasks increases from about 5%, when the statistical information is given in probabilities, to about 25% when it is presented in natural frequencies (McDowell & Jacobs, 2017).

2.1.2. *Visualisations*

Another strategy is to visualise the statistical information of the Bayesian situation (Brase, 2009; Garcia-Retamero & Hoffrage, 2013; Reani et al., 2018) with, for instance, tree diagrams, 2 × 2 tables, double trees, or unit squares (Fig. 2). A broader overview of different visualisations for Bayesian situations is given by Khan et al. (2015) or

[1] The latter two, i.e., P(I|H) and P(I| $\overline{H}$), are notations for conditional probabilities and read as "probability of I, given H″ or "probability of I, given not H″ respectively.

Framing of the Bayesian situation about prenatal screening of down syndrome

Imagine you are working as a gynaecologist. With every pregnant woman you perform a triple-test between the 15th and 18th week of pregnancy in order to detect a possible Down syndrome of the unborn child.

You are currently providing consultation to a 45-year-old woman who has tested positive in the triple-test.

Now, this woman wants to know what this means for her unborn child. Statistics on 45-year-old pregnant women and on the triple-test reveal:

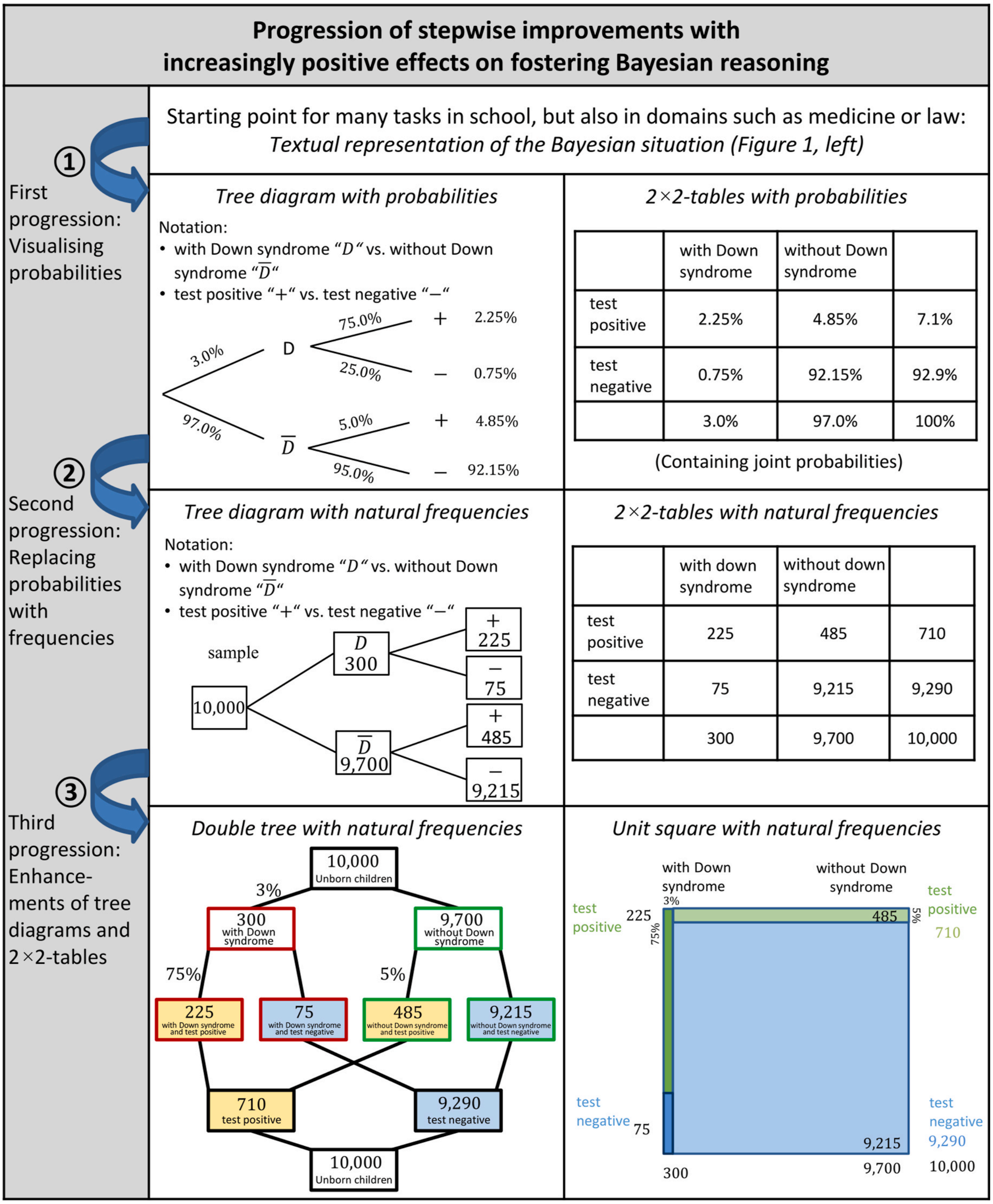
	Given statistical information	
	Probabilities	**Natural frequencies**
Base rate $P(D)$	There is a **3%** probability for a pregnant woman to carry an unborn child *with* Down syndrome (D).	**300** out of **10,000** pregnant women carry an unborn child *with* Down syndrome (D).
True-positive rate $P(+\|D)$	If the pregnant woman carries an unborn child *with* Down syndrome (D), then the probability is **75%** that this woman tests positive (+).	**225** out of the **300** pregnant women carrying an unborn child *with* Down syndrome (D), test positive (+).
False-positive rate $P(+\|\bar{D})$	If the pregnant woman carries an unborn child *without* Down syndrome ($\bar{D}$), then the probability is **5%** that this woman still tests positive (+).	**485** out of the **9,700** pregnant women carrying an unborn child *without* Down syndrome ($\bar{D}$), still test positive (+).
	Questions and answers	
Positive predictive value (PPV)	If a pregnant woman tests positive, what is the probability that she is carrying an unborn child *with* Down syndrome?	Out of the pregnant women who test positive, how many carry an unborn child with Down syndrome?
Possible solution algorithm	$P(D\|+) = \frac{P(D)\cdot P(+\|D)}{P(D)\cdot P(+\|D)+P(\bar{D})\cdot P(+\|\bar{D})}$ $= \frac{\mathbf{0.03}\cdot\mathbf{0.75}}{\mathbf{0.03}\cdot\mathbf{0.75}+(1-\mathbf{0.03})\cdot\mathbf{0.05}} \approx \mathbf{31.69\%}$	**225** out of **710** (= 225 + 485)
Negative predictive value (NPV)	If the pregnant woman tests negative, what is the probability that she is carrying an unborn child *without* Down syndrome?	Out of the pregnant women who test negative, how many carry an unborn child without Down syndrome?
Possible solution algorithm	$P(\bar{D}\|\mp) = \frac{P(\bar{D})\cdot P(\mp\|\bar{D})}{P(\bar{D})\cdot P(\mp\|\bar{D})+P(D)\cdot P(\mp\|D)}$ $= \frac{(1-\mathbf{0.03})\cdot(1-\mathbf{0.05})}{(1-\mathbf{0.03})\cdot(1-\mathbf{0.05})+\mathbf{0.03}\cdot(1-\mathbf{0.75})}$ $\approx \mathbf{99.19\%}$	**9,215** (= 9,700 – 485) out of **9,290** [= 9,215 + (300 – 225)]

Fig. 1. Example of a Bayesian situation based on probabilities (left) and natural frequencies (right) with authentic statistical information (Health Quality Ontario, 2019).

Spiegelhalter et al. (2011). In the following, we focus on commonly used visualisations (i.e., in school and university teaching) and enhancements of these.

When teaching probabilities in school, tree diagrams and 2 × 2 tables filled with probabilities are often utilised (Fig. 2, first progression). This is, for example, evident in many national as well as international curricula and mathematics textbooks. Empirical studies, however, show that such tree diagrams or 2 × 2 tables are only of limited help, yet when probabilities are replaced in the corresponding visualisations with natural frequencies (Fig. 2, second progression), the understanding of the situation can be improved substantially (Binder et al., 2015). Today, visualisations based on natural frequencies can be considered the most promising combination of representational strategies for gaining insight into Bayesian situations (e.g., McDowell & Jacobs, 2017).

Double trees and unit squares (Fig. 2, third progression) are further enhancements of tree diagrams and 2 × 2 tables. Both visualisations can boost performance to about 60%, for instance, compared to only about 30% based on a frequency tree diagram (Böcherer-Linder & Eichler, 2019). A salient advantage of the unit square compared to a 2 × 2 table is that it can display conditional probabilities, which are originally presented in Bayesian situations, i.e., 75% and 5% (Büchter, Steib, et al., 2022). Note that the 2 × 2 table (Fig. 2, above) contains joint


Progression of stepwise improvements with increasingly positive effects on fostering Bayesian reasoning
Starting point for many tasks in school, but also in domains such as medicine or law: Textual representation of the Bayesian situation (Figure 1, left)
① First progression: Visualising probabilities
② Second progression: Replacing probabilities with frequencies
③ Third progression: Enhance-ments of tree diagrams and 2×2-tables
Tree diagram with probabilities
Notation:
• with Down syndrome "D" vs. without Down syndrome "D̄"
• test positive "+" vs. test negative "–"
3.0%
97.0%
D
D̄
75.0%
25.0%
5.0%
95.0%
+ 2.25%
– 0.75%
+ 4.85%
– 92.15%
2×2-tables with probabilities
(Containing joint probabilities)
Tree diagram with natural frequencies
sample
10,000
D 300
D̄ 9,700
+ 225
– 75
+ 485
– 9,215
2×2-tables with natural frequencies
Double tree with natural frequencies
10,000 Unborn children
3%
300 with Down syndrome
9,700 without Down syndrome
75%
5%
225 with Down syndrome and test positive
75 with Down syndrome and test negative
485 without Down syndrome and test positive
9,215 without Down syndrome and test negative
710 test positive
9,290 test negative
10,000 Unborn children
Unit square with natural frequencies
with Down syndrome
without Down syndrome
3%
test positive 225
75%
485
5%
test positive 710
test negative 75
9,215
test negative 9,290
300
9,700
10,000

	with Down syndrome	without Down syndrome	
test positive	2.25%	4.85%	7.1%
test negative	0.75%	92.15%	92.9%
	3.0%	97.0%	100%

	with down syndrome	without down syndrome	
test positive	225	485	710
test negative	75	9,215	9,290
	300	9,700	10,000

Fig. 2. Stepwise evolution of helpful representations of the Bayesian situation given in Fig. 1.

probabilities that are not given at the starting point but first would have to be derived from the typically given conditional probabilities (i.e., true-positive rate and false-positive rate).

It may be argued that other visualisations, e.g., so-called icon arrays[2], are even more promising (Brase, 2008; Gigerenzer et al., 2021). This is admittedly the case if the completed visualisation is already constructed and presented (Böcherer-Linder & Eichler, 2019). However, for the purpose of learning and instruction, it is important that visualisations can easily be sketched by teachers and students, as only in this way is it possible to tackle the starting point where no additional support is provided. Yet, drawing an icon array is very time-consuming in many authentic Bayesian situations (Binder et al., 2015). For example, in the prenatal screening in Fig. 1, a number of at least 10,000 icons would have to be depicted.

In contrast, both double tree and unit square a) display all probabilities given in a Bayesian situation, b) have proven supportive in empirical studies so far, especially when completely filled with natural

[2] An icon array is a visualisation in which icons represent the natural frequencies. Thus, for the situation in Fig. 1, there would be 10,000 icons (e.g., circles) to represent all pregnant women. Some of these icons would be highlighted in a certain way (e.g., colored) to represent the different subsets in the sample.

frequencies and c) can be easily sketched by teachers and students themselves. In short, both visualisations are suitable not only for teaching purposes, but also for a training study aimed at enabling participants to handle Bayesian situations.

2.2. *Explicit training courses*

Providing the representational strategies above (i.e., double tree or unit square based on natural frequencies) helps to improve Bayesian reasoning, but is still not sufficient to reach expert standard (Büchter, Eichler, et al., 2022). Moreover, since in many domains probabilities are usually available without a supportive representation, it is necessary to explicitly train Bayesian reasoning based on this initial situation and to clarify how to deal with the given probabilities for a better understanding.

A number of studies have already implemented training courses about Bayesian reasoning, partially showing quick improvements (for an overview of eleven existing training studies, see Büchter, Eichler, et al., 2022). The aim of the present paper is to combine the following aspects, which were considered only separately in previous training studies:

- *Authentic tasks* for measuring Bayesian reasoning, i.e., probabilities as a starting point (without visualisation) and realistic contexts with genuine statistical information.
- A combination of *both most helpful representational strategies* so far (i.e., visualisations based on natural frequencies).
- Studying *both short- and medium-term effects* (i.e., pre-post-follow-up design).
- *Generalisability* of results, based on a sufficiently *large sample* and the implementation of a *control group* without a training course.
- *A high internal validity* with as many further aspects as possible experimentally controlled.

Authentic tasks (in the above sense) were implemented, for instance, in Hoffrage, Krauss, et al. (2015). Sometimes frequency-based visualisations were implemented in training studies (e.g., Chow & van Haneghan, 2016; Feufel et al., 2023; Ruscio, 2003; Sirota et al., 2015; Steckelberg et al., 2004; Wassner, 2004) showing mixed effects when compared to, e.g., a probability training course (for large effects see e.g., Wassner, 2004: for no effects see e.g., Sirota et al., 2015). Few training studies so far have conducted a pre-post-follow-up design aiming to address both short- and medium-term learning (e.g., Bea, 1995; Sedlmeier & Gigerenzer, 2001; Wassner, 2004). Interestingly, a number of training studies have compared different training courses without implementing a control group without any training (for exceptions, see e.g., Bea, 1995; Sirota et al., 2015; Talboy & Schneider, 2017). However, only the parallel implementation of an additional group without any training guarantees a valid estimation of the "true" effects of training courses. Last but not least, the two visualisations double tree and unit square (see Fig. 2) have not yet been compared to each other in a standardised training study, nor have they been experimentally compared to respective de-compositions, i.e., "probability tree only" (as a school-typical visualisation without natural frequencies) or "natural frequencies only".

2.3. *Influence of prior mathematical achievement*

Since Bayesian reasoning is relevant for a diverse group of people with varying mathematical skills, it seems important to identify the role of individual prerequisites for learning Bayesian reasoning.

In studies without previous instruction or training, numeracy and more general measures of mathematical and cognitive skills can predict performance in Bayesian reasoning (e.g., Brase, 2021; Bruckmaier et al., 2021; Johnson & Tubau, 2015; Sirota et al., 2014).

Furthermore, a range of findings showed that measures of prior (mathematical) achievement influence the learning of statistics (Chance et al., 2022; Kogan & Laursen, 2014) or academic achievement in general (Blömeke, 2009; Hattie, 2009; Schneider & Preckel, 2017). Such studies often use final (mathematics) grades in school as estimates of prior mathematical achievement.

We are not aware of existing research investigating the effect of prior mathematical achievement on the *learning processes* of Bayesian reasoning. However, only knowledge on the interaction of individual prerequisites with different training courses would allow to develop tailored teaching for various groups of learners.

Moreover, as the training courses of our study differ primarily with regard to the visualisation (see below), studying the effect of prior mathematical achievement will provide insights into the varying demands of learning with different visualisations. This seems important as it is recognised that there is no "one-size-fits-all visualisation" (Liu et al., 2020, p. 693). Still, to the best of our knowledge, research about individual differences and visualisations has not yet focused on the learning but rather on the performance with given visualisations (Hall et al., 2022).

2.4. *Research interest and hypotheses*

The present study investigates the short-term (directly after training) and medium-term (after about 8 weeks) learning effects between five computer-based experimental conditions. The special characteristic of the present study is that two competing training courses were constructed "as-optimal-as-possible" while keeping both of them parallel. Both training courses are based on a combination of the two supportive strategies, i.e., to visualise the situation *and* to use natural frequencies ("level-2"). These two training courses are (1) *double tree based on natural frequencies* and (2) *unit square based on natural frequencies* which compete against each other as so-called "level-2 training courses" in the sense of a "betting model" (Verschaffel, 2018). Additionally, two "level-1 training courses" (i.e., with only *one* helpful strategy) are implemented, namely (3) *natural frequencies only* and a (4) *school training* course, based on a probability tree that is typically used in school teaching. Finally, a control group (5) without any training ("level-0") was implemented.

The first research question is, whether the level-1 training courses are in fact more supportive than not receiving any training as in the control group (RQ1) which we hypothesise in H1. In particular, we aim to challenge the effects of the elements which are usually taught in school under controlled conditions. Moreover, we want to analyse whether the level-2 training courses (double tree and unit square; both based on natural frequencies) are in fact more supportive than the level-1 training courses (RQ2), as we expect in H2.

Additional research questions look at which of both level-2 training courses performs better (RQ3) and how the school training course is ranked (RQ4). Concerning all training courses, we are interested which differential effects regarding prior mathematical achievement can be observed (RQ5). RQ3-5 are studied without explicit hypotheses.

3. Material and methods

3.1. *Participants*

Our participants comprised $n = 255$ law and $n = 260$ medical students. About 30% ($n = 162$) of the sample identified as men and 70% ($n = 351$) as women ($n = 2$ participants as other). Age varied from 18 to 35 years ($M = 21.6$; $SD = 2.8$) and the semester of the students ranged from 1 to 20 ($M = 5.5$; $SD = 6.3$). Participation in the study was voluntary; written informed consent was obtained from the participants. All students of both domains received payment (~75$ per person; ~38,625$ in total, funded by the DFG[3]) if they were participating at all three measurement points. The Ethics Commission of the University of Kassel approved the study (zEK-18).

[3] German Research Foundation (DFG: Deutsche Forschungsgemeinschaft).

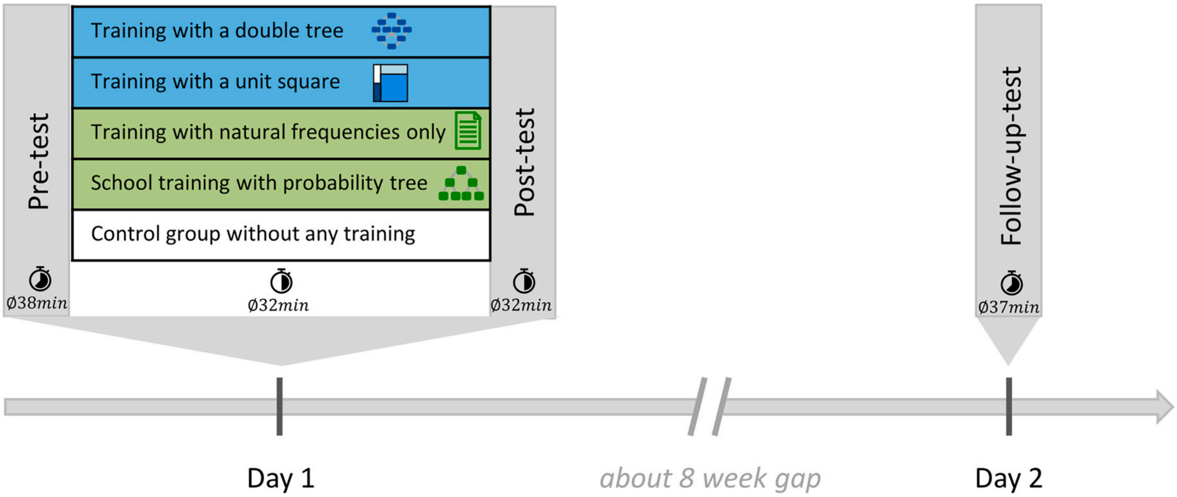

Pre-test
Training with a double tree
Training with a unit square
Training with natural frequencies only
School training with probability tree
Control group without any training
Post-test
Ø38min
Ø32min
Ø32min
Follow-up-test
Ø37min
Day 1
about 8 week gap
Day 2

Fig. 3. Study design (blue: level-2 trainings, green: level-1 trainings, white: level-0 without any training). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.2. Study design

The effectiveness of the different training courses was examined in a pre-post-follow-up design (Fig. 3).

The pre-test, training course and post-test were all carried out on the same day (day 1). The follow-up-test (day 2) was conducted about 8 weeks ($M = 8.3$; $SD = 1$) after day 1 of the study. For the construction of the three tests, see section 3.3 on tasks and materials. The training course took place between pre- and post-test and lasted 32 min on average (input: 22 min; exercise: 10 min).[4]

In each domain, participants were randomly assigned to one of the five different experimental conditions (see section 3.4 for a description of the training courses): *double tree training* (level 2), *unit square training* (level 2), *natural frequencies only training* (level 1), *school training with a probability tree* (level 1) or *control group without a training course* (level 0). No significant differences regarding the covariates (e.g., gender, age, semester) between the five experimental conditions were observed.

3.3. Tasks and materials

All Bayesian situations (for both law and medicine) and the corresponding tasks were constructed under supervision of experts from both domains (e.g., professors of law and professors of medicine) and can be found in the supplementary material A.1 (English translations) and A.2 (original wording in German). A total of 14 Bayesian situations (7 from law and 7 from medicine) were constructed (Table 1).

These tasks are combined in the following way to constitute the pre-, post- and follow-up-tests (also compare Table 1): In the pre-test, four "domain-specific" tasks (= from the domain of the participants) had to be answered (d1, d1*, d2 and d3; with * indicating a question for the NPV). The post-test (five tasks) included two tasks from the pre-test (d1, d1*), two domain-specific tasks not set before (d4, d5) and a "transfer-task" (t1), i.e., from the other domain (e.g., about prenatal screening for law students). The follow-up-test (six tasks) consists of two tasks from the pre-test (d1, d1*), one task from the post-test (d4), two domain-specific tasks not set before (d6, d7) and one transfer-task not set before (t2). Thus, the first presented Bayesian situation and questions about the PPV (d1) and the NPV (d1*) was part of all three measurement points (i.e., d1 and d1* served as anchor items).

Bayesian reasoning was measured by asking for the conditional probability PPV (or NPV). In all Bayesian situations, the three pieces of statistical information (base rate, true-positive rate, and false-positive rate) were given as probabilities without a visualisation. The participants were asked to enter their calculated or estimated PPV (or NPV) as a probability. For each task participants were required to submit a percentage with two decimals. They could not skip any tasks; thus, there are no missing answers for any task.

Prior research in Bayesian reasoning has identified errors that are typically committed by participants when asked to estimate the PPV (Binder et al., 2020; Eichler et al., 2020; Woike et al., 2023). These known errors are summarised for the Bayesian situation regarding prenatal screenings (Fig. 1) in Table 2. In our study, an answer was considered correct if it deviated no more than 0.5% percentage points from the correct value. This interval never included any of the well-known errors and was primarily allowed to accommodate for rounding errors. Additionally, we classified whether the given answer corresponds to one of the known errors.

Prior mathematical achievement was measured by the final mathematics grade in secondary school (with values from 0 to 15, where 15 is the best and 0 the worst grade).

3.4. Training courses

An overview of the different representations of statistical information in the four computerised training courses on Bayesian reasoning is given in Fig. 4. The training courses were constructed based on an elaborated approach for teaching interventions, namely multimedia principles (Mayer, 2009), and the 4C/ID model (Frerejean et al., 2019) including a worked example (Renkl, 2014) on how to calculate the PPV in a Bayesian situation.

In each training course, first, an introduction with technical terms was given (e.g., what is meant with base rate, true-positive rate and false-positive rate). Afterwards, a worked example consisting of three steps was used to explain how to calculate the PPV (see below for details), followed by some practical information (e.g., typical wording of conditional probabilities). After that, an exercise in an authentic Bayesian situation (i.e., d3 from pre-test) followed, in which the participants received individual feedback to their answer. The core part of the training courses are the worked example, which followed the introduction (see the following).

[4] The materials of the training courses and the tests covers also two extensions of conventional Bayesian reasoning (="*calculation*"), namely *covariation*, which requires judging the consequences when input variables change (Büchter et al., 2024; Steib et al., 2023) and *communication*, which refers to explaining the correct result (Böcherer-Linder et al., 2022). However, these additional tasks are not the focus of the present paper.

Table 1
Given information and correct answer to all tasks in the implemented Bayesian situations of the study and usage of each task in the different tests.

Domain	Bayesian situation	Hypothesis	Information	Numerical information (probabilities)			Correct answer		Task		Use of the tasks in the tests					
							PPV	NPV			Pre		Post		Follow-up	
		H	*I*	P(*H*)	P(*I*|*H*)	P(*I*|$\overline{H}$)	P(*H*|*I*)	P($\overline{H}$|$\overline{I}$)	Law	Med	Law	Med	Law	Med	Law	Med
Law	*polygraph*	*Knowledge about a crime*	*hits in a polygraph*	*50%*	*85%*	*10%*	*89.47%*		*d1*	*t1*	*X*		*X*	*X*	*X*	
								85.71%	*d1**	—	*X*		*X*		*X*	
	recidivism	*future recidivism*	*previous conviction*	*41%*	*10%*	*3%*	*69.85%*		*d2*	—	*X*					
	breathalyzer test	*intoxication (alcohol)*	*positive test result in breathalyzer*	*10%*	*93%*	*50%*	*17.13%*		*d3*	—	*X*					
	paternity test	*paternity*	*positive test result in a paternity test*	*5%*	*100%*	*10%*	*34.48%*		*d4*	—			*X*		*X*	
	facial recognition software	*banned football fan*	*facial recognition software sets off an alarm*	*0.5%*	*80%*	*1%*	*28.67%*		*d5*	—			*X*			
	COMPAS algorithm	*future recidivism of a criminal offender*	*COMPAS algorithm sets off an alarm*	*42%*	*70%*	*10%*	*83.52%*		*d6*	*t2*					*X*	*X*
	plagiarism software	*plagiarized work*	*plagiarism software sets off an alarm*	*5%*	*23%*	*2%*	*37.70%*		*d7*	—					*X*	
Med	*Covid antibody test*	*antibodies for Covid-19*	*positive antibody test result*	*6%*	*97%*	*2%*	*75.58%*		*t1*	*d1*		*X*	*X*	*X*		*X*
								99.80%	—	*d1**		*X*		*X*		*X*
	mammography	*breast cancer*	*Mammography positive*	*6%*	*80%*	*10%*	*33.80%*		—	*d2*		*X*				
	Covid self-test	*infection with Covid-19*	*positive Covid-19 self-test result*	*5%*	*96%*	*2%*	*71.64%*		—	*d3*		*X*				
	prenatal screening	*trisomy 21*	*positive triple test result*	*3%*	*75%*	*5%*	*31.69%*		—	*d4*				*X*		*X*
	colon cancer screening	*colon cancer*	*positive hemoccult test result*	*0.5%*	*40%*	*5%*	*3.86%*		—	*d5*				*X*		
	pregnancy test	*pregnant*	*positive pregnancy test result*	*2%*	*99%*	*0.5%*	*80.16%*		*t2*	*d6*					*X*	*X*
	HIV test	*HIV*	*positive HIV test result*	*2%*	*100%*	*0.3%*	*87.18%*		—	*d7*						*X*

Note: d1-d7 denote the tasks in domain-specific (= from the study domain of the participants) Bayesian situations; "*" marks the tasks in which the NPV (negative predictive value) was asked; t1 and t2 denote "transfer-tasks" in a Bayesian situation (i.e., from the other domain); all answers had to be entered with two decimals and are, hence, also listed here with two decimals.

Note that the detailed steps of the worked examples were always based on a "more general" Bayesian situation, structurally equivalent to the Bayesian situation (Fig. 1) but abstracted to a more universal situation with an unspecified piece of evidence or medical test result and their general implication about criminal charges or diseases (see supplementary materials B.1 for the training courses for medicine translated into English, and B.2 and B.3 for the original German training courses for medicine and law, respectively).

The first step of the worked example is to draw the representation of the training course (e.g., the nodes and branches of the double tree without numerical information) and to add the given probabilities to the representation (e.g., add percentages on the branches of the double tree). In the second step, in the training courses with natural frequencies, the frequencies were added to the representation of the training course (e.g., into all nodes in the double tree, see Fig. 4) based on the given probabilities. In this step, the given probabilities are reinterpreted as proportions before translating these into natural frequencies. In the third step of the worked example, the calculation for the PPV with the complete representation is explained in each of the training courses. The three worked examples in training courses with natural frequencies (double tree, unit square, natural frequencies only) are constructed as completely parallel apart from the fact that, in the training "natural frequencies only", no visualisation is drawn.

The school training course deviates from this parallelism, as it is supposed to represent what is typically carried out in school teaching. Here, in the second step no frequencies were added to the representation – as tree diagrams in textbooks in school typically do not display frequencies but probabilities – instead the tree diagram is completed by calculating the joint probabilities and adding them at the end of each path (Fig. 4). Moreover, prior to the worked example, the participants of the school training course revised the addition and multiplication rules for probabilities, as this is relevant for working with the probability tree diagram but not with the other representations.

Thus, in the school training course, the focus was on parallelism to the textbooks in school and not on parallelism to the level-2 training courses. Nevertheless, while adhering to structure and expressions from textbooks, here we also used multimedia principles (Mayer, 2009). With the help of experts (e.g., experienced mathematics teachers) and a thorough textbook analysis, we aimed to create a promising school-typical implementation to give the school training course a fair chance (see supplementary material B.1 to B.3).

3.5. Administration

All students participated in groups (with a maximum number of 38 individuals per group) in a laboratory setting where a computer was provided for each participant. Students worked individually at their own pace in a digital study environment set up in the survey-software *Unipark,* in which Java codes were implemented for all tests and training courses. The digital study environment provided information about the structure and progress of the study, so that no further interaction of the participants was necessary with the researcher during the study. The participants were allowed to take notes on a paper while completing the tests, to allow them to apply the strategies learned in the training courses (i.e., draw a visualisation), with any output handed in after each test. Participants could take a break between the pre-test and the training course (average break time: 10 min) and between the training course and the post-test (average break time: 3 min). The total average time was about 2.5 h for day 1 and about 1 h for day 2.

3.6. Statistical analysis

In our analyses, we use linear mixed regression models (LMMs) to predict the proportion of correct solutions to the Bayesian reasoning tasks at the three measurement points (i.e., pre-, post-, or follow-up-test) based on the different experimental conditions and on the individual

Table 2
Typical errors for estimating the PPV in the example on prenatal screenings for Down syndrome (base rate = 3%; true-positive rate = 75%; false-positive rate = 5%; correct PPV: 31.69%).

Name of the known error	Fisherian	Joint occurrence	Base rate only	Likelihood-subtraction	Likelihood	Pre-Bayes	Evidence only
Error indicated as the PPV = $\mathrm{P}(H\|I)$	$\mathrm{P}(I\|H)$	$\mathrm{P}(H\cap I)$	$\mathrm{P}(H)$	$\mathrm{P}(I\|H) - \mathrm{P}(I\|\overline{H})$	$\frac{\mathrm{P}(H\cap I)}{\mathrm{P}(H\cap \overline{I})}$	$\frac{\mathrm{P}(H)}{\mathrm{P}(I)}$	$\mathrm{P}(I)$
Value in the Bayesian situation about prenatal screening	75%	2.25%	3%	70%	46.39%	42.25%	7.1%

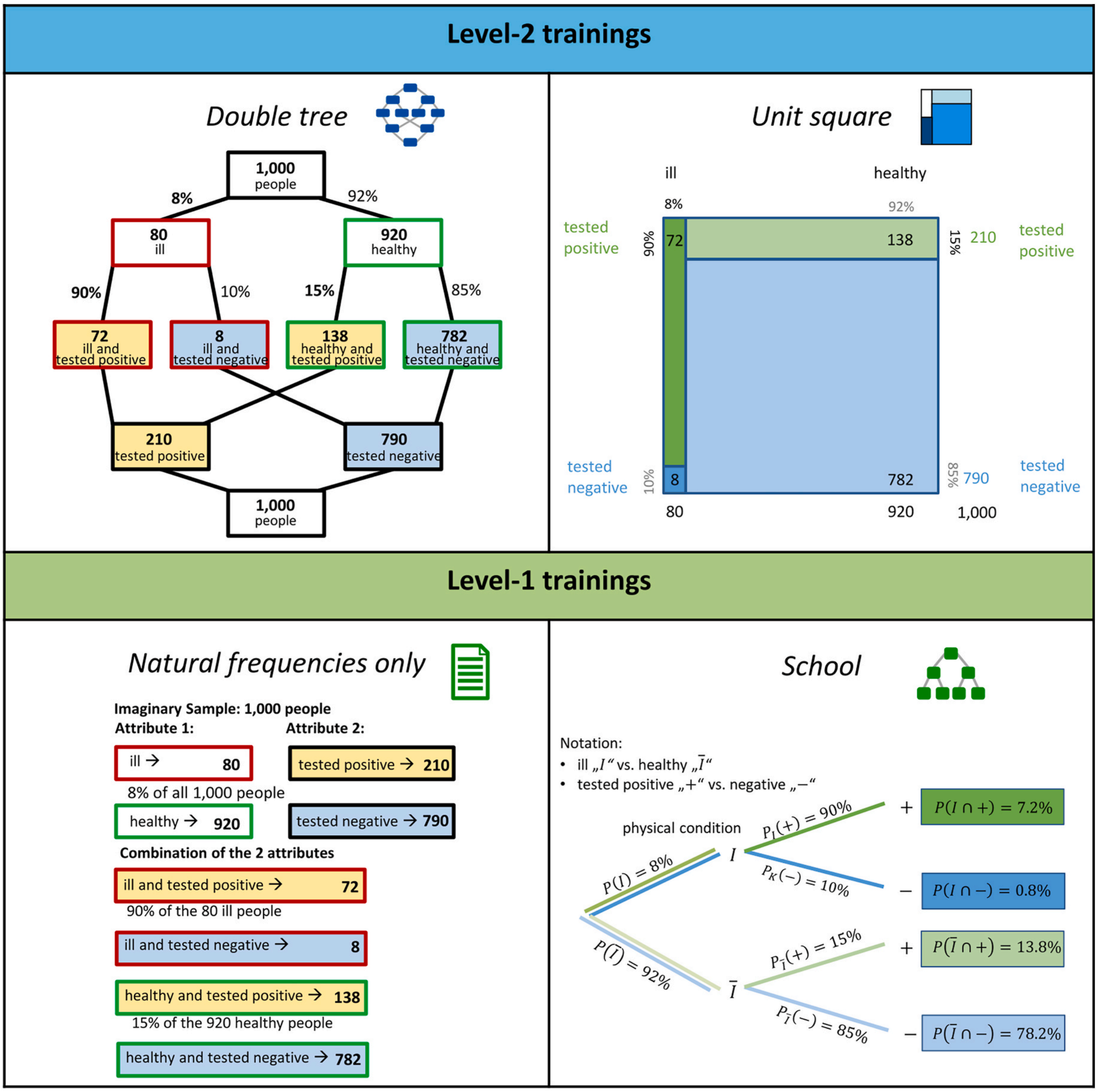


Fig. 4. Representations of statistical information in the four training courses.

prior mathematical achievement (Hilbert et al., 2019). We run two LMMs (see formulae of both models below) to predict the short- and medium-term learning effects separately (LMM1 from pre- to post-test and LMM2 from pre- to follow-up-test). Fixed factors of the model are measurement point (pre-, post- or follow-up-test), the experimental condition, and prior mathematical achievement, as well as the corresponding interactions of these predictors. The reference of the measurement point is the pre-test and the reference for the increases from pre- to post- or to follow-up-test are the increases in the “natural frequencies only” group (in which a medium increase is expected).

The other experimental conditions (DTGroup = double tree training; USGroup = unit square training; SchoolGroup = school training; CONGroup = control group without any training) and measurement points (Post = post-test; FollowUp = follow-up-test) are dummy coded variables in the model (hence, have the value 1 if applicable and 0 otherwise). The estimates for these variables determine a change in the prediction between the dummy-coded categories, for example, between pre-test (coded as 0) and post-test (coded as 1). The only metric (non-dummy-coded) variable is prior mathematical achievement (MA), which was standardised for the model. Therefore, estimated values for MA in the model denote increases or decreases in the predicted proportion of correct solutions to the Bayesian reasoning tasks with a change of MA from the average (coded as 0) to one standard deviation above average (coded as 1).

Random factors can be implemented into mixed models (such as LMMs) as possible sources of errors for non-independent data (Brauer & Curtin, 2018). The (1|ID) and (1|domain) terms in the formulae denote random intercepts for the values nested within the categorical variables 'domain' and 'ID', in order to control for differences between the domains and individual persons in the pre-test.[5]

The models for predicting the proportion of correct solutions in the Bayesian reasoning tasks are given with the following formulae:

$$\begin{aligned} LMM1 : \widehat{y_1} &= \beta_{1.0} + \beta_{1.1}Post + \beta_{1.2}MA + \beta_{1.3}Post \times DTGroup + \beta_{1.4}Post \\ &\times USGroup + \beta_{1.5}Post \times SchoolGroup + \beta_{1.6}Post \times CONGroup \\ &+ \beta_{1.7}Post \times MA + \beta_{1.8}Post \times DTGroup \times MA + \beta_{1.9}Post \\ &\times USGroup \times MA + \beta_{1.10}Post \times SchoolGroup \times MA + \beta_{1.11}Post \\ &\times CONGroup \times MA + (1|domain) + (1|ID) \end{aligned} \quad \text{(equation 2)}$$

$$\begin{aligned} LMM2 : \widehat{y_2} &= \beta_{2.0} + \beta_{2.1}FollowUp + \beta_{2.2}MA + \beta_{2.3}FollowUp \times DTGroup \\ &+ \beta_{2.4}FollowUp \times USGroup + \beta_{2.5}FollowUp \times SchoolGroup \\ &+ \beta_{2.6}FollowUp \times CONGroup + \beta_{2.7}FollowUp \times MA \\ &+ \beta_{2.8}FollowUp \times DTGroup \times MA + \beta_{2.9}FollowUp \times USGroup \\ &\times MA + \beta_{2.10}FollowUp \times SchoolGroup \times MA + \beta_{2.11}FollowUp \\ &\times CONGroup \times MA + (1|domain) + (1|ID) \end{aligned} \quad \text{(equation 3)}$$

For being able to test all hypotheses regarding the order of the different training courses, we also added post-hoc tests: for that we ran the same models but one with the school and another with the unit square training course as reference for the increases from pre- to post- or follow-up-test.

For our analyses we used the statistical software R 4.3.0 (R Core Team, 2016) with the lme4 package (Bates et al., 2012). The resulting p-values were computed with the lmerTest package (Kuznetsova et al., 2017), the resulting pseudo R^2 with the MuMIn Package (Barton, 2023). The data and the R-script can be accessed in the supplementary materials C.1 and C.2, respectively.

4. Results

All $n = 255$ law and $n = 260$ medical students took part in all three measurement points.[6] Therefore, we have no missing data. All groups in each domain are equally distributed ($n = 51$ participants in each of the five groups of law students and $n = 52$ participants in each group of medical students).

Firstly, and before comparing experimental conditions, a substantial overall difference in performance between both domains stands out (Table 3; also see Fig. 5) that was not expected to this degree: medical students clearly outperformed law students. Due to the sufficient reliabilities with high measures of internal consistency regarding pre-test, post-test and follow-up-test (see Cronbach's alphas in Table 3), we use the reliable average scores as indicators of Bayesian reasoning performances below (Fig. 5). Secondly, from a descriptive perspective, these differences remain relatively stable across the three measurement points (thus, the training courses seem to work similarly in both domains). The large differences between the domains might be explained by two possible reasons:

a) Varying difficulty between the tasks used for each domain.
b) Individual differences of the participants between both domains (e.g., regarding the prior mathematical achievement).

Since a) can be excluded considering the similar performances in domain-specific ("d") and transfer ("t") tasks (see Table 3), we follow up on b). There is a large difference between participants of both domains with respect to prior mathematical achievement (MA): law students have an average of 9.7 ($SD = 3.3$) while medical students have an average of 12.7 ($SD = 2.3$) out of 15 points ($t = -12.0$; $p < 0.01$). Differences regarding other covariates (e.g., gender, age, semester) were, however, negligible. Hence, differences between both domains build to a large extent on differences in MA.

Again, the domain was not modelled as a fixed factor because it is not part of the hypotheses or research questions (see above) and we refrained from post-hoc modelling adjustments. Instead of running two different models in both domains as a result of the striking differences, we decided to combine both sub-samples for the models of section 3.6. Since we are especially interested in which training course is appropriate for which prior mathematical achievement, we can obtain more general results for this question by combining both sub-samples. Modelling both domains separately would, in any case, not change our results or conclusions substantially (see supplementary material D). Some domain-specific exceptions will be explained in 4.1.

In Table 4, the effects regarding LMM1 (pre-post) and LMM2 (pre-follow-up) are reported separately. While, in section 4.1, both models are initially discussed without considering MA, in 4.2 the role of MA will be focused for short-term (LMM1) and medium-term effects (LMM2).

4.1. *Learning gains in the five experimental conditions*

In Fig. 5 (lower part, in which the performances for both domains are aggregated), the double tree training course descriptively shows the largest learning gains (short-term and medium-term), while the other three training courses display smaller yet still substantial effects.

In the LMM models, the intercepts of 15% ($\beta_{1.0}$ and $\beta_{2.0}$) stand for the estimated proportion of correct solutions in the pre-test (across all groups, because no pre-test differences between groups were modelled as fixed effects). The regression coefficient for "Post" (or "Follow-Up") represents the estimated increase in the mean proportion of correct solutions from the pre- to post-test (or follow-up-test) in the reference group with the natural frequencies only training:[7] hence, an additional 41% in the post-test (see $\beta_{1.1}$) and an additional 19% in the follow-up-test (see $\beta_{2.1}$; also compared to the pre-test). These increases were also significantly larger than in the control group, as the mean increase from pre- to post-test is estimated to be 30% lower (see $\beta_{1.6}$), and 9% lower (see $\beta_{2.6}$) for the estimated mean increase from pre- to follow-up-test in the control group. Compared to the group with the natural frequencies only training, the short- and medium-term improvements were not significantly larger in the group with the school training, see $\beta_{1.5}$ and $\beta_{2.5}$. Thus, taken together, H1 was supported for both level-1 training courses concerning short-term and medium-term learning.

Moreover, in the group with the double tree training, short- and medium-term improvements were significantly larger than in the group with the natural frequencies only training, as the mean proportion of

[5] Domain was not implemented as a fixed factor because there was no hypothesis or research question about this variable. In our study, domains serve instead for the purpose of mutual validation.

[6] Note: payment was subject to participation in the follow-up-test.

[7] Unlike the previous effect for the intercept, the effect for "Post" only refers to the group with natural frequencies only and not to all groups. This is the case because for "Post" the interactions with the groups are added in the model, for example, "Post × DTGroup".

Table 3
Percentage of correct answers in the five experimental conditions, separated by domain, measurement point and task.

		Double tree		**Unit square**		**Natural frequencies only**		**School**		**Control group**	
Domain		Law §	Med ⚕	Law §	Med ⚕	Law §	Med ⚕	Law §	Med ⚕	Law §	Med ⚕
N		51	52	51	52	51	52	51	52	51	52
	task										
Pre-test	d1	8%	12%	14%	21%	4%	21%	6%	23%	12%	12%
	d1*	8%	17%	12%	25%	4%	25%	2%	15%	6%	10%
	d2	2%	21%	6%	21%	0%	31%	0%	29%	2%	21%
	d3	4%	29%	8%	42%	0%	38%	0%	35%	8%	33%
	$Ø_{domain}$	5%	20%	10%	27%	2%	29%	2%	25%	7%	19%
	$Ø_{all}$	13%		19%		16%		14%		13%	
	$\alpha_{Law} = 0.79$; $\alpha_{Med} = 0.84$; $\alpha_{all} = 0.84$										
Post-test	d4	61%	87%	39%	69%	35%	90%	31%	81%	8%	33%
	d5	49%	81%	29%	83%	24%	77%	24%	69%	6%	29%
	d1	63%	85%	39%	69%	27%	81%	53%	81%	16%	38%
	d1*	51%	83%	37%	75%	20%	79%	25%	87%	14%	37%
	t1	53%	92%	41%	73%	24%	92%	24%	88%	14%	50%
	$Ø_{domain}$	55%	85%	37%	74%	26%	84%	31%	81%	11%	37%
	$Ø_{all}$	70%		56%		55%		57%		24%	
	$\alpha_{Law} = 0.86$; $\alpha_{Med} = 0.84$; $\alpha_{all} = 0.88$										
Follow-up test	d6	24%	52%	18%	33%	4%	46%	12%	44%	14%	23%
	d7	37%	73%	16%	52%	8%	62%	10%	54%	16%	38%
	d1	47%	71%	27%	56%	10%	56%	8%	71%	18%	33%
	d1*	45%	83%	27%	52%	12%	65%	6%	69%	16%	35%
	d4	39%	75%	14%	58%	6%	58%	6%	62%	4%	33%
	t2	37%	85%	18%	60%	8%	69%	14%	77%	16%	44%
	$Ø_{domain}$	38%	74%	20%	52%	8%	59%	9%	63%	14%	34%
	$Ø_{all}$	56%		36%		34%		36%		24%	
	$\alpha_{Law} = 0.92$; $\alpha_{Med} = 0.89$; $\alpha_{all} = 0.92$										

Note: * = tasks, in which the NPV was asked;
d = domain-specific task; t = transfer task;
$Ø_{domain}$ = average percentage of correct answers within each domain
$Ø_{all}$ = average percentage of correct answers across both domains
$\alpha_{Law}/\alpha_{Med}/\alpha_{all}$ = Cronbach's alphas as estimates for internal consistency

correct solutions increased by an additional 16% for the post-test (see $\beta_{1.3}$) and an additional 23% for the follow-up-test (see $\beta_{2.3}$). However, contrary to our hypothesis, in the group with the unit square training, the mean short- and medium-term improvements did not significantly exceed those of the group with the natural frequencies only training (see $\beta_{1.4}$ and $\beta_{2.4}$). Hence, H2 was only supported for the training with a double tree, not for the training with a unit square.

The post-hoc analyses (see Fig. 6) reveal: even though the school training was descriptively better than the training with natural frequencies only, the training with a double tree is still significantly superior to the school training. Likewise, although the unit square training was descriptively inferior to the training with natural frequencies only, it still turned out to be significantly more effective than the control group without any training.

We additionally calculated LMM1 and LMM2 for law and medical students separately (see supplementary material D for the results). The results suggest that the effects reported are indeed comparable between both student groups with a notable exception: the unit square training and the natural frequency training seem to interact with the domain, showing significantly better results of the unit square training than the training with natural frequencies only for the law students, but also inferior results of the unit square training compared to the training with natural frequencies only for the medical students. Moreover, among the medical students, the unit square training and, among the law students, the training with natural frequencies only, do not result in medium-term learning compared to the control group without any training.

The effects reported so far represent the estimates for participants with average prior mathematical achievement, because for these participants the standardised MA equals 0.

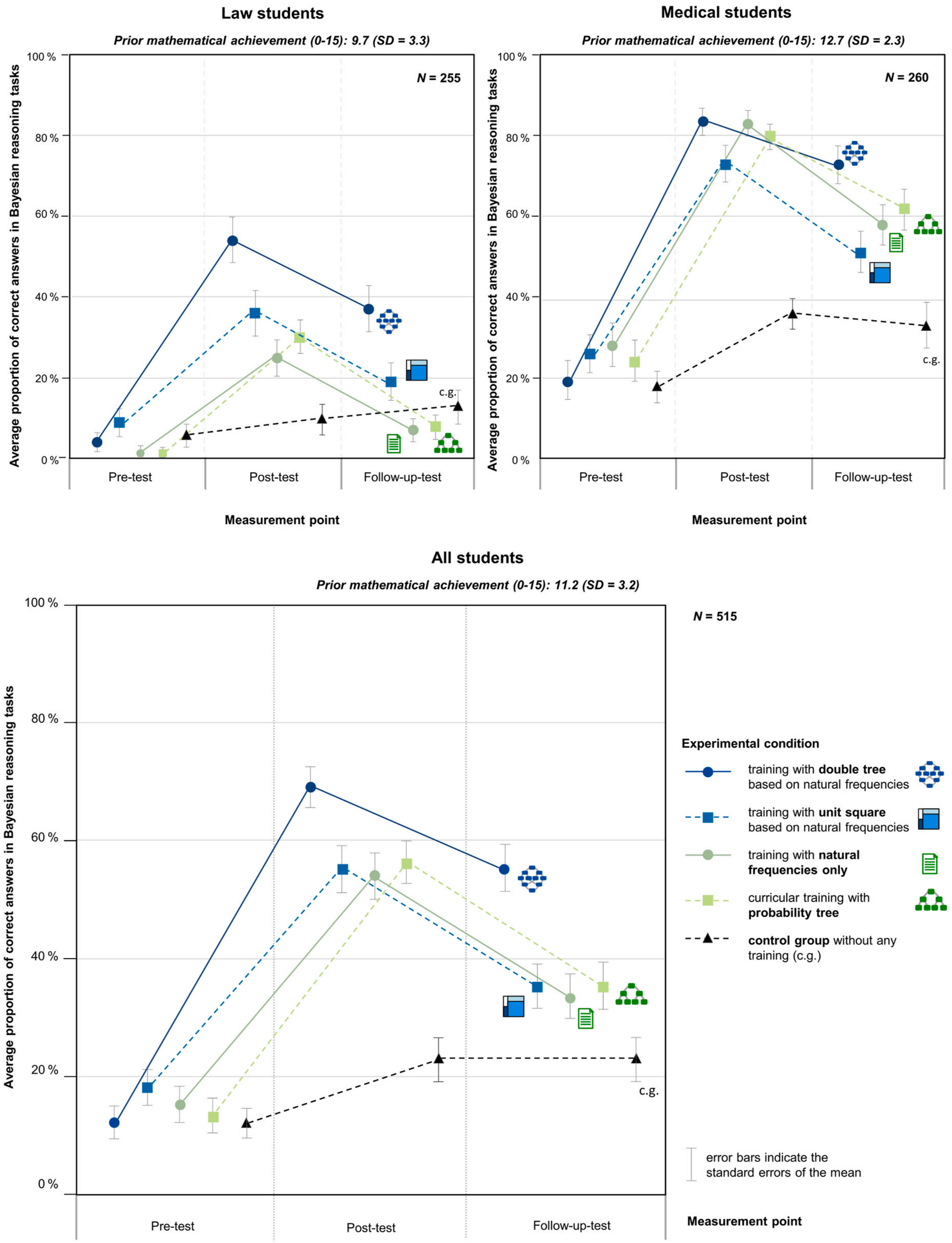


Fig. 5. Average proportion of correct answers per measurement point and experimental condition, separated for law and medical students (above) and across both domains (below).
Note: The varying locations of the five experimental conditions within each measurement point are only implemented for the visibility of the standard errors but do not signify a temporal delay between the five experimental conditions.

4.2. Influence of prior mathematical achievement on learning gains in five different experimental conditions

In the following, we consider the role of MA for short- and medium-term learning in the different experimental conditions. Descriptively, the positive slopes in Fig. 7 suggest that the increases from pre- to post- (or to follow-up-) test (displayed on the y-axis) were higher for students with higher mathematical achievement (displayed on the x-axis) in most groups (i.e., students with higher mathematical achievement learn more). Interestingly, learning with the double tree training seems – at

Table 4
Results of LMM1 and LMM2. Reference of the measurement point is the pre-test and reference for the increases from pre-to post- or follow-up-test is the natural frequencies only training.

LMM1 (short-term effects: Pre-Post)

	$\beta_{1.k}$	$SE_{\beta_{1.k}}$	$t_{\beta_{1.k}}$	p
Intercept ($\beta_{1.0}$)	0.15	0.11	1.3	0.42
***Post* ($\beta_{1.1}$)**	**0.41**	**0.03**	**13.7**	**<0.01**
***MA* ($\beta_{1.2}$)**	**0.05**	**0.01**	**3.52**	**<0.01**
***Post* × *DTGroup* ($\beta_{1.3}$)**	**0.16**	**0.04**	**4.02**	**<0.01**
Post × *USGroup* ($\beta_{1.4}$)	−0.02	0.04	−0.56	0.58
Post × *SchoolGroup* ($\beta_{1.5}$)	0.02	0.04	0.54	0.59
***Post* × *CONGroup* ($\beta_{1.6}$)**	**−0.3**	**0.04**	**−7.67**	**<0.01**
***Post* × *MA* ($\beta_{1.7}$)**	**0.06**	**0.03**	**2.34**	**0.02**
Post × *DTGroup* × *MA* ($\beta_{1.8}$)	−0.06	0.04	−1.52	0.13
Post × *USGroup* × *MA* ($\beta_{1.9}$)	0.04	0.04	1.14	0.26
Post × *SchoolGroup* × *MA* ($\beta_{1.10}$)	0.01	0.04	0.17	0.87
Post × *CONGroup* × *MA* ($\beta_{1.11}$)	−0.05	0.04	−1.29	0.2
$R^2_{Marginal} = 0.32$; $R^2_{Conditional} = 0.62$				

LMM2 (medium-term effects: Pre-Follow-Up)

	$\beta_{2.k}$	$SE_{\beta_{2.k}}$	$t_{\beta_{2.k}}$	p
Intercept ($\beta_{2.0}$)	0.15	0.1	1.42	0.39
***FollowUp* ($\beta_{2.1}$)**	**0.19**	**0.03**	**6.74**	**<0.01**
***MA* ($\beta_{2.2}$)**	**0.05**	**0.01**	**3.71**	**<0.01**
***FollowUp* × *DTGroup* ($\beta_{2.3}$)**	**0.23**	**0.04**	**6.07**	**<0.01**
FollowUp × *USGroup* ($\beta_{2.4}$)	−0.01	0.04	−0.19	0.85
FollowUp × *SchoolGroup* ($\beta_{2.5}$)	0.04	0.04	0.92	0.36
***FollowUp* × *CONGroup* ($\beta_{2.6}$)**	**−0.09**	**0.04**	**−2.36**	**0.02**
***FollowUp* × *MA* ($\beta_{2.7}$)**	**0.06**	**0.03**	**2.21**	**0.03**
FollowUp × *DTGroup* × *MA* ($\beta_{2.8}$)	0.03	0.04	0.89	0.37
FollowUp × *USGroup* × *MA* ($\beta_{2.9}$)	0.02	0.04	0.65	0.52
FollowUp × *SchoolGroup* × *MA* ($\beta_{2.10}$)	0.03	0.04	0.78	0.44
FollowUp × *CONGroup* × *MA* ($\beta_{2.11}$)	−0.05	0.04	−1.42	0.16
$R^2_{Marginal} = 0.19$; $R^2_{Conditional} = 0.62$				

Note: $\beta_{i.k}$ = estimated regression coefficients (unstandardised for $\beta_{i.0}, \beta_{i.1}, \beta_{i.3} - \beta_{i.6}$; semi-standardised for $\beta_{i.7} - \beta_{i.11}$; standardised for $\beta_{i.2}$)
$SE_{\beta_{i.k}}$ = standard error of the estimated regression coefficients.
$t_{\beta_{i.k}}$ = t-value of each estimated regression coefficient.
p = probability for committing a type-I error.
$R^2_{Marginal}$ = variance explained by fixed effects.
$R^2_{Conditional}$ = variance explained by both fixed and random effects.

least regarding short-term learning – to have been almost independent from previous mathematical achievement, as the slope is close to 0.

In both LMM models, the regression coefficients of 5% for "MA" ($\beta_{1.2}$ and $\beta_{2.2}$) stand for mean increase in the proportion of correct answers in the pre-test across all experimental conditions, if MA is one standard deviation above average. We are particularly interested in the influence of MA on the improvements from pre- to post- or follow-up-test, which can be seen in the interactions with MA. Hence, the p-values for the regression coefficients for "Post × MA" and "FollowUp × MA" ($\beta_{1.7}$ and $\beta_{2.7}$) show that, in the group with the natural frequencies only training[8], the influence of MA on the performance was significantly larger in the post- and follow-up-test than in the pre-test, suggesting that the short- and medium-term learning effects of participants with higher MA were larger than those of participants with lower MA. The effects of the other interactions with MA imply that the influence of MA on the short- and medium-term learning was not significantly different in the other groups compared to the natural frequency group.

Nevertheless, the post-hoc analyses reveal that the influence of MA was significantly larger in the unit square group than in the double tree group for the short-term learning ($\beta = -0.1; SE_\beta = 0.04; t_\beta = -2.47; p = 0.01$) but not for the medium-term learning (for the estimated models, see end of section 3.6). This implies that the level-2 training courses particularly differed regarding the influence of MA on the short-term learning results: in the double tree group, participants with high and low MA may have learned equally well in the short-term (Fig. 7), but training with the unit square seems to have been particularly suitable for participants with higher MA.

4.3. Analysis of typical errors

As we are particularly interested in comparing which training course is most effective for enabling students to provide correct answers, we limited our inferential analyses to the proportion of correct answers (as reported above). However, we also checked whether participation in the training courses helped to reduce typically known errors regarding the estimation of the PPV (Table 2). These analyses revealed that all errors are reduced in the post-test through participation in all of the training courses and also in the follow-up-test (though to a lesser extent). Furthermore, these reductions in errors do not essentially differ between the training courses, with one exception: regarding the changes from pre- to post-test the Fisherian strategy is reduced to a greater degree in the three training groups with a visualisation (i.e., training with a double tree, training with a unit square and school training) than in the training with natural frequencies only. The reduction of the Fisherian strategy in the training group with natural frequencies only seems comparable to the reduction of this strategy in the control group without any training course.

4.4. Qualitative validation of the results

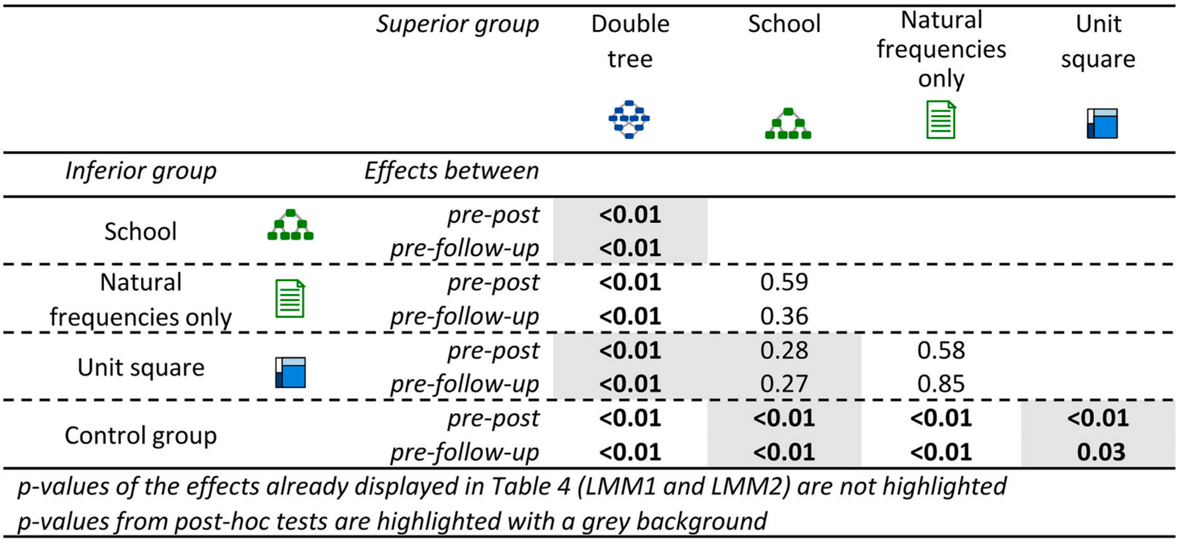
The results regarding the training with a unit square do not confirm our hypothesis (H2). In order to provide explanations for these surprising findings, we carried out a further qualitative validation of our results by analysing the notes of the participants. For that, we coded (in each task of the pre-, post- and follow-up-test) if and what visualisation was sketched as a representation of the Bayesian situation in the task. In Fig. 8, we display the results of this analysis. In the pre-test, mostly (single) tree-diagrams (not double trees) and 2 × 2-tables were sketched by the participants, if a visualisation was used at all. In the post- and follow-up-test, the representation of the training course was used in all training groups. However, it is notable, that in the school training group the proportion of students using the visualisation from the training (i.e., a probability tree-diagram) was the highest. This is particularly remarkable in the follow-up-test. Moreover, it emerges clearly that the proportion of students sketching the unit square in the post- or follow-up-test is smaller than that of students sketching the double tree. The same is true for the students with the natural frequencies only training.

5. Discussion

In the present study, the effectiveness of four different training courses on Bayesian reasoning (double tree, unit square, natural frequencies only, probability tree) was investigated in a pre-post-follow-up design with $n = 255$ students of law and $n = 260$ students of medicine. The results show short- and medium-term learning effects with all training courses and a superiority of the training course with double tree over the others. Furthermore, students with higher prior mathematical achievement (MA) profited more from the training courses than those with low previous MA (only short-term learning with the double tree training was independent from MA).

Our results are informative and valuable for the education of future experts with a need for Bayesian reasoning. However, we consider these results also as relevant for teaching probabilities in school. This claim of

[8] Unlike the previous effect for the influence of MA in the pre-test, the effect for "Post × MA" and "FollowUp × MA" only refers to the reference group and not to all groups. This is the case, because for "Post × MA" and "Follow-Up × MA" the interactions with the groups are added in the model, for example, "Post × DTGroup × MA".



	Superior group	Double tree	School	Natural frequencies only	Unit square
Inferior group	*Effects between*				
School	*pre-post*	**<0.01**			
	pre-follow-up	**<0.01**			
Natural frequencies only	*pre-post*	**<0.01**	0.59		
	pre-follow-up	**<0.01**	0.36		
Unit square	*pre-post*	**<0.01**	0.28	0.58	
	pre-follow-up	**<0.01**	0.27	0.85	
Control group	*pre-post*	**<0.01**	**<0.01**	**<0.01**	**<0.01**
	pre-follow-up	**<0.01**	**<0.01**	**<0.01**	**0.03**

p-values of the effects already displayed in Table 4 (LMM1 and LMM2) are not highlighted
p-values from post-hoc tests are highlighted with a grey background

Fig. 6. Contrasts between the experimental conditions in descriptive order of their ranks with respect to short- and medium-term effects.

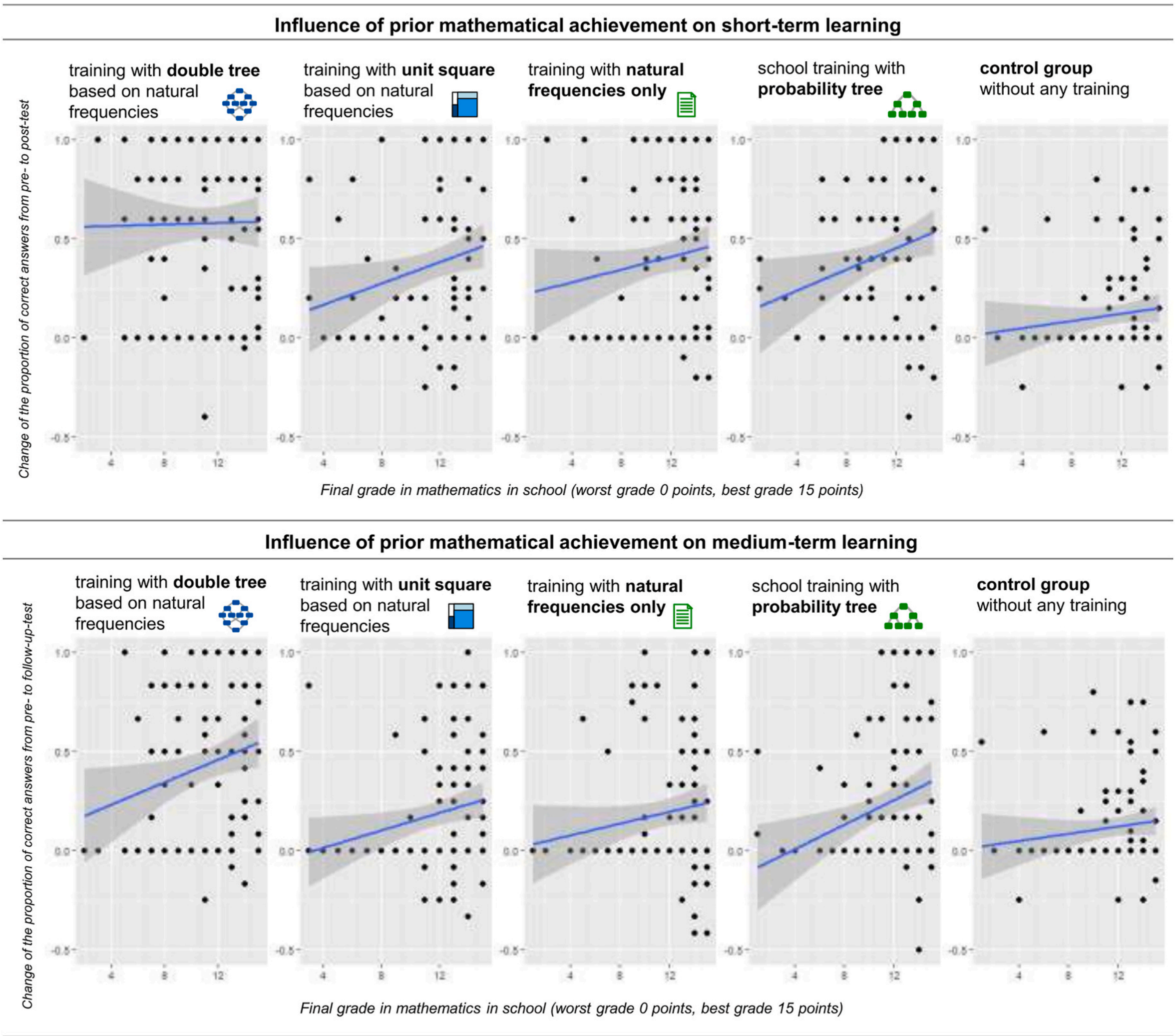


Fig. 7. Scatterplots with slopes indicating the influence of MA (0–15 points) on the increase from pre- to post-test or pre- to follow-up-test (difference of average), separated by the different experimental conditions.

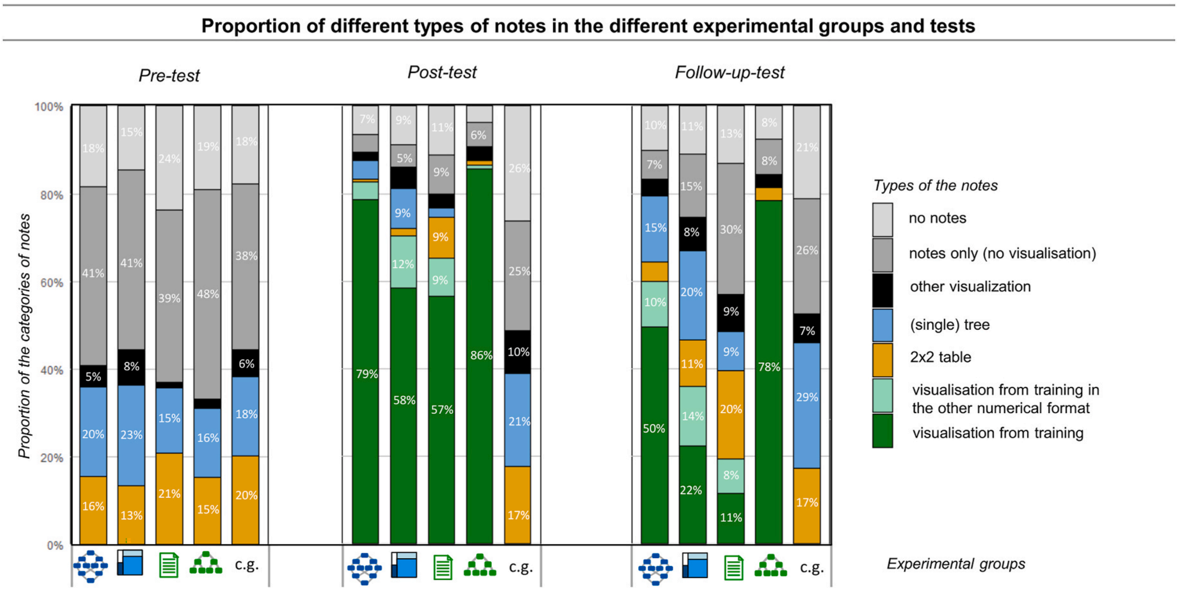


Fig. 8. Proportion of different types of notes in each test and experimental condition.
Note: the experimental conditions from left to right: training with a double tree, training with a unit square, training with natural frequencies only, training with a probability tree, control group without any training; percentages of the proportions of the different categories are only displayed if the respective proportion is 5% or larger.

transferability from tertiary to secondary education of the results rests on the results of the meta-analysis by McDowell and Jacobs (2017), in which it has been shown that experts (e.g., physicians) and non-experts generally do not differ regarding their performance in Bayesian reasoning, and also do not differ in their advantage gained from the supportive representation including the beneficial effect of natural frequencies. Hence, even though our sample may deviate from upper secondary students with regard to expertise in the domains of medicine and law, there is no reason to assume that the learning of Bayesian reasoning should differ substantially between our sample and students in school.

5.1. Learning Bayesian reasoning with different training strategies

Students with the natural frequencies only training showed better short- and medium-term learning results than those without any training. This is consistent with previous literature, showing that training courses with natural frequencies were identified as a successful strategy (Chow & van Haneghan, 2016; Feufel et al., 2023; Sedlmeier & Gigerenzer, 2001; for an exception, see Sirota et al., 2014).

As expected, when natural frequencies are combined with a visualisation in a level-2 training course (specifically with a double tree), learning is stronger in both the short- and medium-term compared to a level-1 training course with natural frequencies only. This is also consistent with previous studies on fostering Bayesian reasoning without instruction, showing a superiority of the frequency double tree over natural frequencies only (Binder et al., 2020). Hence, our results show that the combination of both strategies (visualisation and natural frequencies) is also more effective for instruction. Moreover, apart from comparisons with the training natural frequencies only, the double tree training is also superior to learning with a probability tree, which is the standard method for teaching at school. Overall, as the training course with a double tree based on natural frequencies stands out compared to all other training courses, it seems appropriate for future research to examine its effects for younger students.

However, the results regarding two conditions were unexpected: namely, the training with the unit square and the school training with the probability tree. The results with the school training course may seem surprising as, in previous training studies, instruction with natural frequencies often led to greater improvements than probabilities, particularly for medium-term learning (Sedlmeier & Gigerenzer, 2001). Contrarily, we observed similar short- and medium-term learning in the group with the school and with the natural frequencies only training. This may be attributed to the fact that the school training used a visualisation, namely, the probability tree (unlike the natural frequencies only training), or to the careful multimedia design of the probability tree and supervision of the design by experienced teachers. Hence, visualisations with probabilities (in a convincing design) may be supportive for *learning*, even though they are not supportive *without instruction* (Binder et al., 2015).

Additionally, the qualitative results of the notes are informative: while the ability to construct a probability tree diagram was obtained from post- to follow-up-test, the performance significantly decreased from post- to follow-up-test. Contrarily, in the group with natural frequencies only, both, the proportion of natural frequency notes as well as the performance decreased from post- to follow-up-test. This implies that the probability tree diagram itself may be easy to construct (possibly based on its familiarity) but it is not a helpful representation as itself, which can be interpreted as a replication of prior studies regarding the support of the probability tree diagram (Binder et al., 2015). Yet, the scaffolding of natural frequencies may remain constant as long as the ability to construct them is still given. Consequently, it should be checked whether the comparable results of school and natural frequencies training can in fact be replicated for *long*-term learning as well (e.g., about one year after instruction).

However, the learning results regarding the unit square are the most surprising. Increases after participation in the unit square training do not differ from both level-1 training groups, concerning short- and medium-term. This contradicts various prior findings. Firstly, without instruction, a unit square based on natural frequencies outperforms natural frequencies only (Tsai et al., 2011). Secondly, without instruction, no differences have been observed between a unit square and a double tree (both based on natural frequencies) for Bayesian reasoning (Böcherer-Linder & Eichler, 2019). Thirdly, previous studies with a corresponding paper-pencil training course show increases from a performance of about 10% to as much as 80%–90% in the post-test (Eichler, Gehrke, Böcherer-Linder, & Vogel, 2019).

One reason for the unexpected results with the unit square training could be that natural frequencies were not used ideally in this training (due to the requirement of parallelism of both level-2 training courses): The first step was to draw the (area-proportional) structure based on the given probabilities. This means that the construction of the unfamiliar feature (area-proportionality) rested on probabilities and not on natural frequencies.[9] It could be that using the concept of natural frequencies while structuring the area proportionality would be easier, as done in the paper-pencil training course in the study of Eichler, Gehrke, Böcherer-Linder, & Vogel (2019).

Our results generally raise the question of whether the strategies identified as helpful *without instruction* (e.g., McDowell & Jacobs, 2017) are equally supportive for the *instruction* of Bayesian reasoning. Deviations from our expectations may be based on the varying amount of provided scaffolding by the different visualisations. In typical studies without instruction, visualisations were directly given in the presentation of the Bayesian situation. By contrast, in our training study, a visualisation was not displayed during the tests but only during the training course. Therefore, our participants had to construct the visualisations themselves to access the potential scaffolding. Yet, the challenge of construction may vary for the different visualisations: the probability tree is known from school and, therefore, likely to be the easiest to construct and remember. This is also evident in the additional qualitative analyses as, even eight weeks after the training course, 78% percent of the students were still able to construct the probability tree. The construction may be somewhat harder for the double tree, as unfamiliar elements have to be remembered (i.e., frequencies in the nodes), even though it builds on familiar elements from the tree diagram (i.e., nodes connected by branches). Indeed, a substantial 79% of participants was able to sketch the double tree with natural frequencies in the post-test (slightly fewer than the 86% with the probability tree), yet in the follow-up-test this structure seems harder to remember as only 50% still drew the double tree with natural frequencies (and 10% now drew a double tree with probabilities).

However, constructing the unit square is likely to be the hardest visualisation to construct and remember, as integrating the area-proportionality is a completely unfamiliar feature of the visualisation. This can also be seen in the notes of the corresponding level-2 training groups for both the post-test (79% sketched a double tree vs. 58% a unit square) and follow-up-test (50% sketched a double tree vs. 22% a unit square). Furthermore, the area-proportionality of the unit square might also have been particularly challenging in the present study, as many of the authentic tasks had very small base rates (e.g., 0.1%), which are not possible to draw true to scale. This was different in prior studies with the unit square (Eichler, Gehrke, Böcherer-Linder, & Vogel, 2019). Therefore, the results may imply that either the support of the unit square is limited to situations with higher base rates or that additional training is necessary how to handle Bayesian situations with small base rates in the unit square (e.g., by only approximating the proportions).

These observations may be particularly relevant, as the training study by Feufel et al. (2023) showed that Bayesian reasoning after a training course still differed regarding the numerical format of the given information (i.e., natural frequencies vs. probabilities). The same may be true for effects of different visualisations that are provided simultaneously to textual information in a task. Hence, the possible variation in the challenge of constructing the different visualisations (and thus variations in their scaffolding) may partly explain the surprising results. These effects may have even been strengthened by the fact that actively constructing a visualisation was found to increase performance in one study (Cosmides & Tooby, 1996).

Finally, the qualitative analyses of the error strategies provide some interesting results with regards to the Fisherian strategy that seems to be better avoided when learning with a visualisation (i.e., double tree, unit square or probability tree) than with natural frequencies only. This is a relevant finding, as the Fisherian strategy has dramatic consequences since it leads to drastically overestimating the PPV. For instance, the confusion of the PPV with the true-positive rate, as documented among others by HIV consultants in assuming absolute certainty about positive HIV-test results was based on a high true-positive rate of 99,7% (Prinz et al., 2015). Thus, learning to visualise Bayesian situations may indeed help physicians and counsellors to avoid misdiagnosis.

5.2. *Influence of prior mathematical achievement on learning Bayesian reasoning*

As discussed by Blömeke (2009), prior achievement is a broad measure which comprises both cognitive as well as motivational variables. Accordingly, prior mathematical achievement (MA) can be interpreted differently, for example, as mathematical skills or prior knowledge. Our results show that MA strongly influences Bayesian reasoning in the pre-test already and to a large extent explains differences between law and medical students. This is in line with previous findings on Bayesian reasoning in studies without instruction (Brase, 2021). Our results are unique, however, in showing that also the *learning* of Bayesian reasoning is influenced by MA (higher MA was associated with more learning gains). Thus, the learning of Bayesian reasoning with the strategies used in our study seems to be affected by the “Matthew effect” (Stanovich, 2009). This effect was already studied in other areas of mathematics learning (Kollar et al., 2014) and is likely due to the fact that higher levels of prior knowledge may lead to easier integration of new information into existing knowledge structures.

Interestingly, this influence of MA on learning differs between both level-2 training courses (for short-term learning): a double tree is equally supportive for all participants, in contrast to the unit square, for which higher MA is associated with more learning. One implication would be that learners with low prior achievement may depend particularly on the scaffolding which possibly remains stronger for the double tree than for the unit square (see above). Another explanation could be that different levels of mathematical skills are required for understanding different representations, and the unit square may be a representation which requires high mathematical skills. A further implication is based on differences between both training courses: the double tree training more strongly builds on natural frequencies than the unit square training. Previous results on the dependence of dealing with natural frequencies on mathematical abilities are mixed (Chapman & Liu, 2009; Galesic et al., 2009), at least concerning cross-sectional purely representational studies. However, our results may indicate that natural frequencies are equally supportive for all people, as long as they are combined with an adequate visualisation (i.e., double tree). These insights can help to tailor training courses for Bayesian reasoning to the students' needs.

Our results may also have implications about learning with visualisations in general. Not only in Bayesian reasoning but also in other areas of mathematics education, visualisations are established as a medium to foster understanding (Duval, 2006; Presmeg, 1986; Schoenherr & Schukajlow, 2024). In an up-to-date review on empirical studies about visualisations in mathematics education, however, individual differences are not mentioned as a current area of research (Schoenherr & Schukajlow, 2024). Similarly, in a review by the same authors focusing specifically on interventions with visualisations (Schoenherr & Schukajlow, 2023) individual differences are also not mentioned as characteristics with respect to the corresponding effectiveness.

However, the results of our study imply that the causal relationships between understanding and visualisations are complex, intertwined with individual differences and, more importantly, vary between different visualisations. This resonates with prior research in showing

[9] Note that the size of the nodes and branches in a double tree and probability tree are unaffected by the concrete probability. Therefore, entering the probabilities first may not have been as challenging in the other training courses.

that instruction with a specific visualisation depends on students' prior achievement (Lee et al., 2018). Thus, future research should study, for example, which aspects of a visualisation cause difficulties for some but not others and what kind of previous knowledge and skills are necessary for being able to learn with different visualisations. In other research fields, e.g., computational sciences and psychology, more attention has already been given to individual differences regarding the understanding of visualisations (Cohen & Hegarty, 2007; Liu et al., 2020; Ziemkiewicz et al., 2012).

6. Limitations

Our training courses are tailored to situations where only probabilities without any scaffolding are given, and therefore supportive representations have to be created from scratch. Thus, our results are limited to those representations which seem suitable for these kinds of situations without comparison to other representations (e.g., icon arrays) which may be more supportive in situations where the scaffolding is already provided.

Also, the results are limited to elementary Bayesian situations, i.e., Bayesian situations with binary hypothesis and binary information (Zhu & Gigerenzer, 2006). It needs to be checked whether the results can also be replicated for Bayesian situations characterised by multiple hypotheses (e.g., differentiation between different types of trisomy), multiple test results (e.g., positive, negative or unclear test result) or several pieces of information (e.g., various tests with positive or negative outcomes, Krauss et al., 1999).

Additionally, algorithm-based learning cannot be excluded; however, the similar solution rates for the questions about the NPV (whose calculation was not explained in neither the worked example nor the exercises) suggest that the strategies were adopted conceptually.

Furthermore, we are aware that the implications of our results are limited to the high standardisation with which the computerised training courses were designed. While this allowed a systematic comparison between instructional methods, it needs to be checked whether the results can be transferred to instruction in a less standardised teaching context, for example, in schools.

Finally, the presented results are limited to the learning effects on conventional Bayesian reasoning. In the larger context of our study, further aspects of an extended framework of Bayesian reasoning (http://bayesianreasoning.de/en/br_trainbayes_en.html) including additional competencies such as covariation (i.e., judging the effects of changing input parameters in the Bayesian situation, Büchter et al., 2024, Steib et al., 2023) and communication of the correct result (Böcherer-Linder et al., 2022) were also addressed. It is interesting how the four training conditions affect both the ability for covariation and communication.

7. Conclusion

A training study in a pre-post-follow-up design on the learning of Bayesian reasoning was carried out with five experimental conditions: two optimally designed (level-2) training courses (each with a visualisation – double tree or unit square – based on natural frequencies), two other promising (level-1) training courses, and a control group (level-0). Moreover, the dependence of the effects of the training strategy on prior mathematical achievement was studied.

It was shown that Bayesian reasoning in authentic Bayesian situations can be improved for short- as well as medium-term learning with all implemented training courses. Participants with high prior mathematical achievement learned more than those with average or lower prior achievement. The training course with the double tree stands out, as it shows better learning success than all other training courses. Consequently, using a double tree with natural frequencies is more supportive than a strategy often implemented in national curricula and more common in textbooks, such as the probability tree. From our perspective, this raises the question of why textbooks, national curricula and teaching practices still seem to cling to representations which are less supportive for learning Bayesian reasoning.

CRediT authorship contribution statement

Nicole Steib: Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Theresa Büchter:** Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Andreas Eichler:** Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Karin Binder:** Writing – review & editing, Visualization, Methodology, Formal analysis, Conceptualization. **Stefan Krauss:** Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Katharina Böcherer-Linder:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Markus Vogel:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Sven Hilbert:** Writing – review & editing, Methodology, Formal analysis.

Research data

All data as well as the R-script for the data analyses are available.

Conflict of interest

We have no conflict of interest to declare.

Funding sources

The research leading to these results received funding from Deutsche Forschungsgemeinschaft (DFG) under the Grants EIC773/4-1 and KR2032/6-1.

Declaration of competing interest

The authors declare no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2024.102032.

References

Barton, K.. MuMIn: Multi-Model Inference. R package version 1.47.5. https://CRAN.R-project.org/package=MuMIn.

Batanero, C., & Álvarez-Arroyo, R. (2024). Teaching and learning of probability. *ZDM – Mathematics Education, 56*(1), 5–17. https://doi.org/10.1007/s11858-023-01511-5

Batanero, C., Chernoff, E. J., Engel, J., Lee, H. S., & Sánchez, E. (2016). Research on teaching and learning probability. In C. Batanero, E. J. Chernoff, J. Engel, H. Lee, & E. Sanchez (Eds.), *ICME-13 topical surveys. Research on teaching and learning probability* (pp. 1–33). Imprint: Springer International Publishing. https://doi.org/10.1007/978-3-319-31625-3_1.

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes. R package version, 0*, 999999–1000000 [Computer software].

Bea, W. (1995). Stochastisches Denken: Analysen aus kognitionspsychologischer und didaktischer Perspektive. *Psychologie des Entscheidungsverhaltens und des Konfliktes, 6*. Lang.

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. 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*, 750. https://doi.org/10.3389/fpsyg.2020.00750

Blömeke, S. (2009). Ausbildungs- und Berufserfolg im Lehramtsstudium im Vergleich zum Diplom-Studium – Zur prognostischen Validität kognitiver und psycho-motivationaler Auswahlkriterien. *Zeitschrift für Erziehungswissenschaft, 12*(1), 82–110. https://doi.org/10.1007/s11618-008-0044-0

Böcherer-Linder, K., Binder, K., Büchter, T., Eichler, A., Krauss, S., Steib, N., & Vogel, M. (2022). Communicating conditional probabilities in medical practice. In S. A. Peters,

L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Proceedings of the 11.th international conference on teaching statistics*. *IASE*.

Böcherer-Linder, K., & Eichler, A. (2019). How to improve performance in bayesian inference tasks: A comparison of five visualizations. *Frontiers in Psychology, 10*, 267. https://doi.org/10.3389/fpsyg.2019.00267

Borovcnik, M. (2016). Probabilistic thinking and probability literacy in the context of risk Pensamento probabilístico e alfabetização em probabilidade no contexto do risco. *Educação Matemática Pesquisa Revista do Programa de Estudos Pós-Graduados em Educação Matemática, 18*(3). https://revistas.pucsp.br/index.php/emp/article/view/31495.

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review, 15*(2), 284–289. https://doi.org/10.3758/PBR.15.2.284

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

Brase, G. L. (2021). Which cognitive individual differences predict good bayesian reasoning? Concurrent comparisons of underlying abilities. *Memory & Cognition, 49* (2), 235–248. https://doi.org/10.3758/s13421-020-01087-5

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods, 23*(3), 389–411. https://doi.org/10.1037/met0000159

Bruckmaier, G., Krauss, S., Binder, K., Hilbert, S., & Brunner, M. (2021). Tversky and kahneman's cognitive illusions: Who can solve them, and why? *Frontiers in Psychology, 12*, Article 584689. https://doi.org/10.3389/fpsyg.2021.584689

Büchter, T., Eichler, A., Böcherer-Linder, K., Vogel, M., Binder, K., Krauss, S., & Steib, N. (2024). Covariational reasoning in Bayesian situations. *Educational Studies in Mathematics*, 1–25. https://doi.org/10.1007/s10649-023-10274-5

Büchter, T., Eichler, A., Steib, N., Binder, K., Böcherer-Linder, K., Krauss, S., & Vogel, M. (2022). How to train novices in bayesian reasoning. *Mathematics, 10*(9), 1558. https://doi.org/10.3390/math10091558

Büchter, T., Steib, N., Böcherer-Linder, K., Eichler, A., Krauss, S., Binder, K., & Vogel, M. (2022). Designing visualizations for bayesian problems according to multimedia principles. *Education Scienes, 12*(11), 739. https://doi.org/10.3390/educsci12110739

Burril, G. (2020). Statistical literacy and quantitative reasoning: Rethinking the curriculum. In P. Arnold (Ed.), *New skills in the changing world of statistics education proceedings of the roundtable conference of the international association for statistical education*.

Burrill, G., & Pfannkuch, M. (2024). Emerging trends in statistics education. *ZDM – Mathematics Education, 56*(1), 19–29. https://doi.org/10.1007/s11858-023-01501-7

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., & Leader, S. (2022). Student performance in curricula centered on simulation-based inference. *Statistics Education Research Journal, 21*(3), 4. https://doi.org/10.52041/serj.v21i3.6

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making, 4*(1), 34–40.

Chow, A. F., & van Haneghan, J. P. (2016). Transfer of solutions to conditional probability problems: Effects of example problem format, solution format, and problem context. *Educational Studies in Mathematics, 93*(1), 67–85. https://doi.org/10.1007/s10649-016-9691-x

Cohen, C. A., & Hegarty, M. (2007). Individual differences in use of external visualisations to perform an internal visualisation task. *Applied Cognitive Psychology, 21*(6), 701–711. https://doi.org/10.1002/acp.1344

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*(1), 1–73. https://doi.org/10.1016/0010-0277(95)00664-8

Duval, R. (2006). A cognitive analysis of problems of comprehension in a learning of mathematics. *Educational Studies in Mathematics, 61*(1–2), 103–131. https://doi.org/10.1007/s10649-006-0400-z

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

Ellis, K. M., & Brase, G. L. (2015). Communicating HIV results to low-risk individuals: Still hazy after all these years. *Current HIV Research, 13*(5), 381–390. https://doi.org/10.2174/1570162x13666150511125629

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), 75. https://doi.org/10.1186/s12909-023-04025-6

Frerejean, J., Merriënboer, J. J. G., Kirschner, P. A., Roex, A., Aertgeerts, B., & Marcellis, M. (2019). Designing instruction for complex learning: 4c/id in higher education. *European Journal of Education, 54*(4), 513–524. https://doi.org/10.1111/ejed.12363

Gal, I., & Geiger, V. (2022). Welcome to the era of vague news: A study of the demands of statistical and mathematical products in the COVID-19 pandemic media. *Educational Studies in Mathematics, 111*(1), 5–28. https://doi.org/10.1007/s10649-022-10151-7

Galesic, M., & Garcia-Retamero, R. (2010). Statistical numeracy for health: A cross-cultural comparison with probabilistic national samples. *Archives of Internal Medicine, 170*(5), 462–468. https://doi.org/10.1001/archinternmed.2009.481

Galesic, M., Gigerenzer, G., & Straubinger, N. (2009). Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Medical Decision Making: An International Journal of the Society for Medical Decision Making, 29* (3), 368–371. https://doi.org/10.1177/0272989X08329463

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: A Journal of the American Psychological Society, 8*(2), 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*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Gigerenzer, G., Multmeier, J., Föhring, A., & Wegwarth, O. (2021). Do children have Bayesian intuitions? *Journal of Experimental Psychology: General, 150*(6), 1041–1070. https://doi.org/10.1037/xge0000979

Goodie, A. S., & Fantino, E. (1996). Learning to commit or avoid the base-rate error. *Nature, 380*(6571), 247–249. https://doi.org/10.1038/380247a0

Hall, K. W., Kouroupis, A., Bezerianos, A., Szafir, D. A., & Collins, C. (2022). Professional differences: A comparative study of visualization task performance and spatial ability across disciplines. *IEEE Transactions on Visualization and Computer Graphics, 28*(1), 654–664. https://doi.org/10.1109/TVCG.2021.3114805

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Health Quality Ontario. (2019). Noninvasive prenatal testing for trisomies 21, 18, and 13. *Sex Chromosome Aneuploidies, and Microdeletions: A Health Technology Assessment, 4*, 1–166. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6395059/pdf/ohtas-19-1.pdf.

Hilbert, S., Stadler, M., Lindl, A., Naumann, F., & Bühner, M. (2019). Analyzing longitudinal intervention studies with linear mixed models. *TPM - Testing, Psychometrics, Methodology in Applied Psychology, 26*. Article 1 http://www.tpmap.org/wp-content/uploads/2019/03/26.1.6.pdf.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine, 73*(5), 538–540. https://doi.org/10.1097/00001888-199805000-00024

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

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. https://doi.org/10.3389/fpsyg.2015.01473

Johnson, E. D., & Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Frontiers in Psychology, 6*, 938. https://doi.org/10.3389/fpsyg.2015.00938

Kahneman, D., & Tversky, A. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases (1* (pp. 153–163). Cambridge Univ. Press. https://doi.org/10.21236/ada099501.

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. https://doi.org/10.1016/j.ijhcs.2015.07.001

Kleiter, G. D. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer (Ed.), *Recent Research in Psychology. Contributions to mathematical psychology, psychometrics, and methodology* (pp. 375–388). Springer. https://doi.org/10.1007/978-1-4612-4308-3_27.

Kogan, M., & Laursen, S. L. (2014). Assessing long-term effects of inquiry-based learning: A case study from college mathematics. *Innovative Higher Education, 39*(3), 183–199. https://doi.org/10.1007/s10755-013-9269-9

Kollar, I., Ufer, S., Reichersdorfer, E., Vogel, F., Fischer, F., & Reiss, K. (2014). Effects of collaboration scripts and heuristic worked examples on the acquisition of mathematical argumentation skills of teacher students with different levels of prior achievement. *Learning and Instruction, 32*, 22–36. https://doi.org/10.1016/j.learninstruc.2014.01.003

Krauss, S., Martignon, L., & Hoffrage, U. (1999). Simplifying bayesian inference: The general case. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 165–179). Springer US. https://doi.org/10.1007/978-1-4615-4813-3_11. Imprint: Springer.

Krauss, S., Weber, P., Binder, K., & Bruckmaier, G. (2020). Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit – Forschungsdesiderate und einige Antworten. *Journal für Mathematik-Didaktik, 41*(2), 485–521. https://doi.org/10.1007/s13138-019-00156-w

Kurzenhäuser, S., & Hoffrage, U. (2009). Teaching Bayesian Reasoning: An evaluation of a classroom tutorial for medical students. *Medical Teacher, 24*(5), 516–521.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lee, C.-Y., Lei, K. H., Chen, M.-J., Tso, T.-Y., & Chen, I.-P. (2018). Enhancing understanding through the use of structured representations. *Eurasia Journal of Mathematics, Science and Technology Education, 14*(5). https://doi.org/10.29333/ejmste/85424

Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics, 43*, 147–163.

Liu, Z., Crouser, R. J., & Ottley, A. (2020). Survey on individual differences in visualization. *STAR - State of the Art Report, 39*(3). https://arxiv.org/pdf/2002.07950.

Mayer, R. E. (2009). *Multimedia learning* (2). Cambridge Univ. Press. https://doi.org/10.1017/CBO9780511811678

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

Presmeg, N. (1986). Visualisation in high school mathematics. *For the Learning of Mathematics, 6*(3), 42–46. https://www.jstor.org/stable/40247826.

Prinz, R., Feufel, M. A., Gigerenzer, G., & Wegwarth, O. (2015). What counselors tell low-risk clients about HIV test performance. *Current HIV Research, 13*(5), 369–380. https://doi.org/10.2174/1570162x13666150511125200

R Core Team. (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing [Computer software].

Radermacher, W. J. (2022). Statistical awareness promoting a data culture. *Statistical Journal of the IAOS, 38*(2), 453–461. https://doi.org/10.3233/SJI-220956

Reani, M., Davies, A., Peek, N., & Jay, C. (2018). How do people use information presentation to make decisions in Bayesian reasoning tasks? *International Journal of Human-Computer Studies, 111*, 62–77. https://doi.org/10.1016/j.ijhcs.2017.11.004

Renkl, A. (2014). The worked examples principle in multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbooks in psychology. The Cambridge handbook of multimedia learning* (2nd ed., pp. 391–412). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.020.

Reyna, V. F., & Brainerd, C. J. (2007). The importance of mathematics in health and human judgment: Numeracy, risk communication, and medical deicison making. *Learning and Individual Differences, 17*(2), 147–159. https://doi.org/10.1016/j.lindif.2007.03.010

Roberts, C. D., Stough, L. D., & Parrish, L. H. (2002). The role of genetic counseling in the elective termination of pregnancies involving fetuses with disabilities. *The Journal of Special Education, 36*(1), 48–55. https://doi.org/10.1177/00224669020360010501

Ruscio, J. (2003). Comparing Bayes's theorem to frequency-based approaches to teaching Bayesian reasoning. *Teaching of Psychology, 30*(3), 325–328.

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin, 143*(6), 565–600. https://doi.org/10.1037/bul0000098

Schoenherr, J., & Schukajlow, S. (2023). Characterizing external visualization interventions: A systematic literature review. In M. Ayalon, B. Koichu, R. Leikin, L. Rubel, & M. Tabach (Eds.), *Proceedings of the 46th conference of the international group for the psychology of mathematics education, 4* pp. 163–170).

Schoenherr, J., & Schukajlow, S. (2024). Characterizing external visualization in mathematics education research: A scoping review. *ZDM – Mathematics Education, 56* (1), 73–85. https://doi.org/10.1007/s11858-023-01494-3

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130*(3), 380–400. https://doi.org/10.1037//0096-3445.130.3.380

Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict bayesian reasoning. *Psychonomic Bulletin & Review, 21*(1), 198–204. https://doi.org/10.3758/s13423-013-0464-6

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. *Quarterly Journal of Experimental Psychology, 68*(1), 1–9. https://doi.org/10.1080/17470218.2014.972420, 2006.

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science (New York, N.Y.), 333*(6048), 1393–1400. https://doi.org/10.1126/science.1191181

Stanovich, K. E. (2009). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Journal of Education, 189*(1–2), 23–55. https://doi.org/10.1177/0022057409189001-204

Steckelberg, A., Balgenorth, A., Berger, J., & Mühlhauser, I. (2004). Explaining computation of predictive values: 2 x 2 table versus frequency tree. A randomized controlled trial ISRCTN74278823. *BMC Medical Education, 4*, 13. https://doi.org/10.1186/1472-6920-4-13

Steib, N., Krauss, S., Binder, K., Büchter, T., Böcherer-Linder, K., Eichler, A., & Vogel, M. (2023). Measuring people's covariational reasoning in Bayesian situations. *Frontiers in Psychology, 14*, 1184370. https://doi.org/10.3389/fpsyg.2023.1184370

Stine, G. J. (1996). *Acquired immune deficiency syndrome: Biological, medical, social, and legal issues.* Prentice Hall.

Talboy, A. N., & Schneider, S. L. (2017). Improving accuracy on bayesian inference problems using a brief tutorial. *Journal of Behavioral Decision Making, 30*(2), 373–388. https://doi.org/10.1002/bdm.1949

Tsai, J., Miller, S., & Kirlik, A. (2011). Interactive visualizations to improve bayesian reasoning. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting, 55*(1), 385–389. https://doi.org/10.1177/1071181311551079

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science (New York, N.Y.), 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Veaux, R. D. de, Velleman, P. F., & Bock, D. E. (2012). *Intro stats (3. International ed., technology update).* Pearson/Addison-Wesley.

Verschaffel, L. (2018). *Intervention research in mathematics education.* Borken: GDM-Nachwuchskonferenz.

Wassner, C. (2004). *Förderung Bayesianischen Denkens: Kognitionspsychologische Grundlagen und didaktische Analysen.* Franzbecker. Dissertation https://kobra.uni-kassel.de/handle/123456789/2006092214705.

Wegwarth, O., & Gigerenzer, G. (2013). Less is more: Overdiagnosis and overtreatment: Evaluation of what physicians tell their patients about screening harms. *JAMA Internal Medicine, 173*(22), 2086–2087. https://doi.org/10.1001/jamainternmed.2013.10363

West, L. M., & Brase, G. L. (2023). Improving patient understanding of prenatal screening tests: Using naturally sampled frequencies, pictures, and accounting for individual differences. *PEC Innovation, 3*, Article 100197. https://doi.org/10.1016/j.pecinn.2023.100197

Woike, J. K., Hertwig, R., & Gigerenzer, G. (2023). Heterogeneity of rules in bayesian reasoning: A toolbox analysis. *Cognitive Psychology, 143*, Article 101564. https://doi.org/10.1016/j.cogpsych.2023.101564

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

Ziemkiewicz, C., Ottley, A., Crouser, R. J., Chauncey, K., Su, S. L., & Chang, R. (2012). Understanding visualization by understanding individual users. *IEEE Computer Graphics and Applications, 32*(6), 88–94. https://doi.org/10.1109/MCG.2012.120

Eichler, A., Gehrke, C., Böcherer-Linder, K., & Vogel, M. (2019). A training in visualizing statistical data with a unit square. Eleventh Congress of the European Society for Research in Mathematics Education (No. 6). Freudenthal Group; Freudenthal Institute; ERME. https://hal.science/hal-02435232/.