RISTAL 4 / 2021

Research in Subject-matter Teaching and Learning

Volume 4 – Special Issue edited by Timo Leuders & Katharina Loibl

Citation:

Loibl, K., Leuders, T. (2021). Modeling Teachers' Diagnostic Judgments by Bayesian Reasoning and Approximative Heuristics [Special Issue]. *RISTAL, 4*, 88–108.

DOI: https://doi.org/10.23770/

ISSN 2616-7697

RISTAL. Research in Subject-matter Teaching and Learning Special Issue (2021), pp. 88–108

Modeling Teachers' Diagnostic Judgments by Bayesian Reasoning and Approximative Heuristics

Katharina Loibl & Timo Leuders

Abstract

The diagnostic judgments teachers make can be regarded as inferences from manifest observable evidence on students' behavior (e.g., responses to a task) to their latent traits (e.g., misconceptions). The judgment process can be modeled by Bayesian reasoning. We use this framework to analyze the situation of teachers' diagnostic judgments on students' potential misconceptions based on students' responses. Humans typically deviate from normative Bayesian reasoning and apply heuristic strategies, often by only partially processing the available information (e.g., neglecting base rates). From the perspective of ecological rationality, such heuristics possibly constitute viable cognitive strategies for assessing student errors. We investigate the adequacy of a cognitively plausible heuristic strategy, which amounts to approximating the average probability information on prior hypotheses (base rates of student misconceptions) and evidence (student errors). With a computational simulation, we compare this strategy to optimal Bayesian reasoning and to information-neglecting strategies. We interpret the resulting accuracy within the ecology of informal student assessment.

Keywords

diagnostic judgment, Bayesian reasoning, heuristic, computational simulation

Acknowledgments

This work was funded by the Ministry of Science, Research and Arts of Baden-Wuerttemberg within the Research Training Group "Diagnostic Competences of Teachers" (DiaKom).

1 Introduction

A diagnostic judgment of a teacher can be regarded as an inference from manifest observable evidence on a student's behavior to his or her thinking or latent traits. In order to draw such an inference a teacher needs specific knowledge about students' thinking (e.g., typical misconceptions) on the one hand and about students' behavior that may indicate such a misconception (e.g., typical errors) on the other hand.

This description is quite broad and needs a specification with respect to the subject, the content, and student thinking (cf. editorial of this special issue). For example, in mathematics, students who have the frequent misconception that the fractional part of a decimal number refers to a natural number (Moloney & Stacey, 1997), consistently give wrong answers to comparison problems, such as $4.8 < 4.63$ as they compare 8 to 63. However, an uncertainty remains, since even students with this misconception may occasionally solve a task correctly. Also, students without a misconception may occasionally (i.e., with a low probability) give a wrong answer. Furthermore, students can demonstrate the error for a different reason, e.g., by omitting the decimal point

DOI: https://doi.org/10.23770/

altogether: 48 < 463. Such more differentiated situations of judgments based on errors that may have multiple causes are discussed and investigated in Leuders and Loibl (2020).

The process of judging students' solutions can be seen from the larger perspective of teachers' diagnostic competence, and adds a new perspective to the recent research in this area. Diagnostic competence is used as an umbrella term for the knowledge, beliefs, and skills of teachers which enable them to adequately assess or predict student performance (Herppich et al., 2018; Loibl, Leuders, & Dörfler, 2020). Many studies investigate the influence of teacher knowledge or types of diagnostic judgments on the accuracy (Südkamp, Kaiser, & Möller, 2012) by correlational analyses. Recently, there has been an uptick in studies interested in explaining the genesis of diagnostic judgments by drawing on cognitive models of teacher thinking (Krolak-Schwerdt, Pitten-Cate, & Hölstermann, 2018; Loibl et al., 2020). This information-processing perspective raises new and interesting questions: In which way are cues in the diagnostic situation perceived and processed? Can accuracy also be achieved by heuristic processes? How exactly do individuals deal with uncertainty? Uncertainty can be regarded as constitutive element in diagnostic situations that require inference from manifest information to latent traits or processes. Therefore, diagnostic judgments are sometimes considered as a process of reducing uncertainty (Heitzmann et al., 2019).

The research presented in this paper focuses on the role of uncertainty in diagnostic situations and on heuristics in teachers' diagnostic thinking[1]. It has its origin in the attempt to connect research on Bayesian reasoning as a model for decision making in uncertainty to research on diagnostic judgments. Many other facets of diagnostic competence (e.g., the structure of teacher knowledge, the influence of stereotypes) are excluded from this paper.

The example above on judging misconceptions demonstrates that the connection between latent trait and manifest behavior is inherently uncertain. Furthermore, the direction of the inference is from the manifest evidence (the error) to its assumed cause (a possible misconception). Hence, the result of a diagnostic judgment is rather a set of hypotheses about the observed student with varying plausibility than an unequivocal classification of the student. One can easily find similar examples in different topics and other domains: In science class, teachers observe students' explanations for phenomena and infer their conceptual understanding or misconceptions (Hoppe et al., 2020). In language education, teachers regularly infer language skills from the students' errors in verbal utterances (Witzigmann et al., 2021 in this issue).

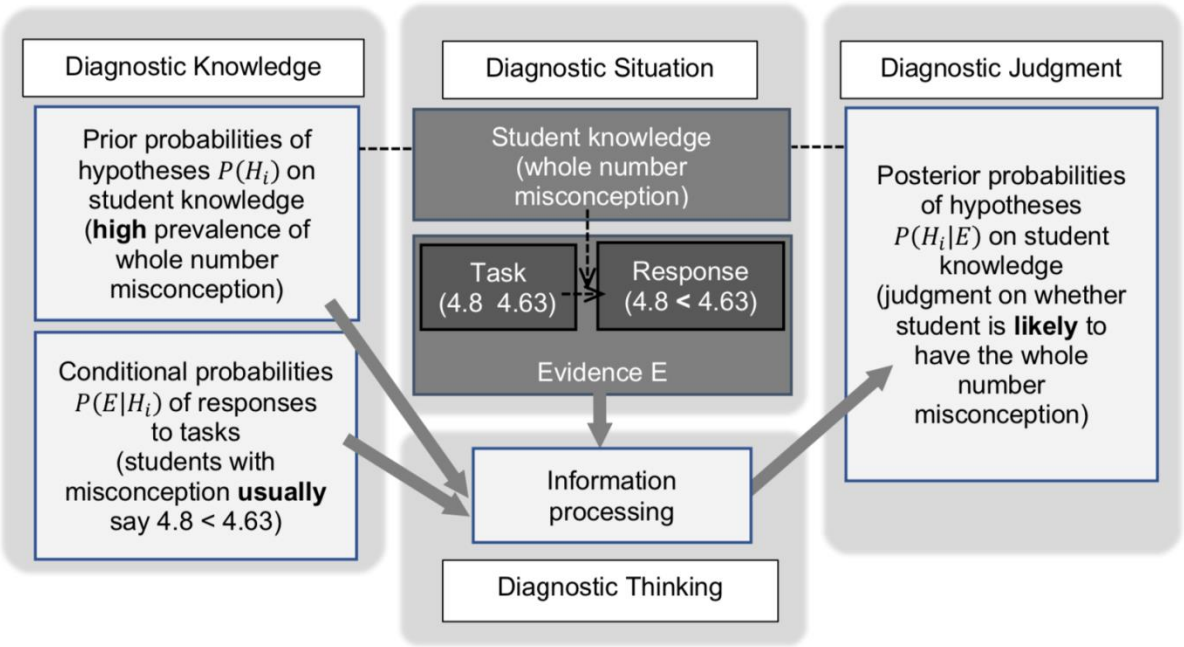
[1] Note that the computational analysis without elaboration on the context of diagnostic judgments of teachers has been included in a conference proceeding in the field of cognitive science, not education (Loibl & Leuders, 2020).

DOI: https://doi.org/10.23770/

The optimal strategy for revising assumptions on the plausibility of different hypotheses after processing new evidence can be described within probability theory as Bayesian reasoning. Within this approach, the plausibility of competing assumptions is described by prior probabilities of the hypotheses $P(H_i)$ (in our example the prevalence of a misconception), the conditional probabilities of evidence by likelihoods of evidence $P(E|H_i)$ (i.e., the probability of a specific error given a misconception), and the revised assumptions by posterior probabilities of hypotheses $P(H_i|E)$.

The structure of this updating process in the context of a teacher's diagnostic judgment on student knowledge is displayed in Figure 1: In order to update the probabilities of the hypotheses (from $P(H_i)$ to $P(H_i|E)$), the teacher processes his or her diagnostic knowledge (i.e., prior probabilities and conditional probabilities) as well as the information provided in the diagnostic situation (i.e., the evidence). Uncertainty plays a major role in this updating process: Students do not respond consistently (cf. conditional probabilities) and different student knowledge may lead to the same responses (ambiguity/ limited diagnosticity).

Fig. 1 (cf. Leuders & Loibl, 2020): The structure of a teacher's diagnostic judgment based on knowledge, evidence, and information processing and the role of uncertainty (with the example of judging a student's misconceptions based on a task response).



Of course, for a teacher these pieces of knowledge and information are usually not explicitly represented by numbers, but only by qualitative and subjective estimations in his or her mind. Any assumed process of Bayesian reasoning therefore also relies on processing such information in a qualitative, non-numerical or fuzzy way.

Bayesian update is a general normative model of decision-making (Mandel, 2014). A judgment based on a Bayesian update strategy (BUS) can be described by the following

DOI: https://doi.org/10.23770/

formula: The revised probabilities of the hypotheses are proportional to the prior probabilities and to the conditional probabilities of the evidence (likelihoods):

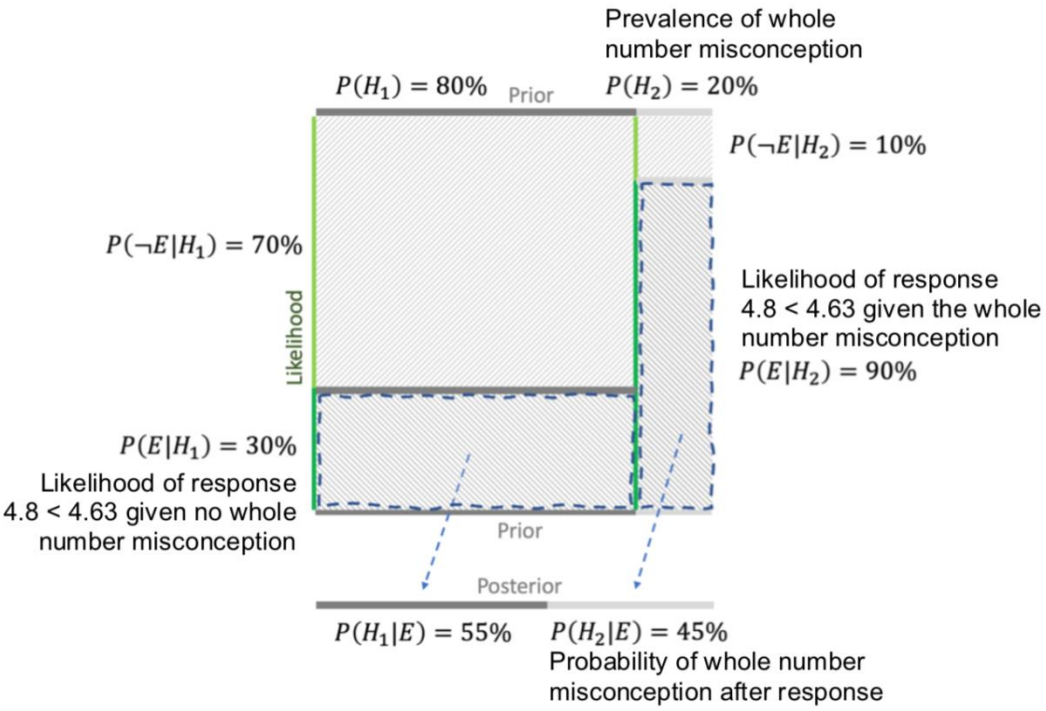
$$P(H_i|E) \propto P(H_i) \cdot P(E|H_i) \qquad \text{(Bayesian update, BUS)} \qquad (1)$$

The diagram in Figure 2 illustrates the updating of the probability of two mutually exclusive hypotheses H_i following the BUS: What is the plausibility of the hypothesis H_2 that a student has the whole number misconception after a response to a task (4.8 < 4.63) in comparison to hypothesis H_1 that he or she does not have this misconception? Given the base rate of the misconception $P(H_i) = 20\%$ and the likelihoods of the student response ($P(E|H_1) = 30\%, P(E|H_2) = 90\%$), after the response is observed the prior probability $P(H_2) = 20\%$ increases to the posterior probability

$$P(H_2|E) = \frac{90\% \cdot 20\%}{30\% \cdot 80\% + 90\% \cdot 20\%} \approx \frac{20\%}{45\%} \approx 45\%.$$

This updated probability corresponds to the ratio of the right dashed area to the total dashed area in Figure 2.

Fig. 2*:* An example of Bayesian reasoning when judging the plausibility of the hypothesis H_2 that a student has the whole number misconception after a response to a task (e.g., 4.8 < 4.63).



Research has demonstrated that humans' capacity to process information on probabilities in Bayesian reasoning is limited, which results in sub-optimal heuristics such as base-rate neglect (e.g., Kahneman & Tversky, 1996). While often such biased

DOI: https://doi.org/10.23770/

strategies are interpreted as a limitation of human thinking in probabilities (Kahneman & Tversky, 1996), one could also ask whether suboptimal heuristics may be regarded as adequate and effective reasoning strategies in certain situations (ecological rationality: Simon, 1955; Gigerenzer & Goldstein, 1996). For instance, Sundh (2019) showed that for calculations with joint probabilities an averaging heuristic was adequate in certain constellations.

In our study, we focus on a situation that has not been studied in the light of ecological rationality before: single-cue judgments on multiple hypotheses with complete but noisy probability information (priors and likelihoods) in the context of teachers inferences on students' misconceptions based on their responses to tasks (evidence). The noisy estimates of the probabilities (cf. Sundh, 2019) make exact calculations unfeasible. From a cognitive perspective, it seems plausible that teachers' mental models of their intuitive estimates on probabilities are analog non-numerical representations (Khemlani, Lotstein, & Johnson-Laird, 2015). Khemlani et al. proposed a computational model assuming primitive analog representations for noisy probabilities and implemented intuitive strategies on processing these non-numerical probabilities in the model. Juslin, Nilsson, & Winman (2009) modeled complex types of reasoning with noisy probabilities (including Bayesian updating, see below). Both computational models were validated with human data.

Against this background, we propose a cognitively plausible heuristic strategy (simpler than Juslin et al., 2009), which amounts to approximately averaging the probability information on prior hypotheses and evidence (APES). We explore the accuracy and ecological rationality of APES for Bayesian reasoning in the context of teacher judgments with noisy estimates of non-numerical probabilities. In two computational studies, we analyze the relative accuracy of judgments based on APES in a situation where the judgment corresponds to the hypotheses that a student does or does not have a misconception as well as to the situation where the teacher decides between three possible misconceptions. Finally, we discuss the scope and ecological rationality of APES by framing it in the context of teacher judgments. While our example is based on misconceptions in decimal comparisons, the computational simulation can be interpreted in any context (i.e., any topic of any subject) in which multiple mutual exclusive hypotheses are realistic.

Our approach owes much to cognitive science, where modeling of human thinking and computational exploration of the implications of the chosen model is a common approach. Since subject specific education research also has a psychological branch that focuses on cognitive processes (of students and teachers), we assume that this approach can contribute a valuable theoretical understanding of possible fine-grained cognitive mechanisms underlying teacher thinking and activity.

2 State of Research on Heuristics in Bayesian Reasoning

DOI: https://doi.org/10.23770/

As outlined above, Bayesian reasoning requires the combination of multiple probabilities in a multiplicative way (e.g., 90% · 20%). However, multiplying probabilities is not intuitive (20% of 90%) and therefore it is cognitively demanding (Sundh, 2019). It therefore does not seem plausible that teachers rely on Bayesian reasoning in their everyday judgments. Unsurprisingly, research (in other contexts) shows that humans often fail to apply the Bayes rule correctly, even when strongly supported (Gigerenzer & Hoffrage, 1995; Weber, Binder, & Kraus, 2018). More specifically, research has identified often-applied heuristics that lead to biased decisions (e.g., base-rate neglect, Kahneman & Tversky, 1996).

In a systematic analysis on the types of update strategies in the context of numerical Bayes reasoning tasks, Cohen and Staub (2015) showed that most participants' strategies amount to not making use of all sources of information: Most participants estimated the posterior probability based on only one of the multiple provided probabilities or by computing a weighted sum of several, but not all probabilities. In their study, most participants only processed the evidence (*evidence only, EOS*, cf. representative thinking: Zhu & Gigerenzer, 2006; Fisherian: Gigerenzer & Hoffrage, 1995; evidence only: Zhu & Gigerenzer, 2006; inverse fallacy: Villejoubert & Mandel, 2002; likelihood subtraction: Gigerenzer & Hoffrage, 1995 for variants of strategies that only focus the evidence). Other participants only took the prior probabilities (*priors only, POS*, cf. base rate only: Gigerenzer & Hoffrage, 1995; also called conservatism: Edwards, 1968; Zhu & Gigerenzer, 2006).

$$P(H_i|E) \propto P(E|H_i) \qquad \text{(Evidence only, EOS)} \quad (2)$$

$$P(H_i|E) \propto P(H_i) \qquad \text{(Prior only, POS)} \quad (3)$$

The reasoning strategies related to (2) and (3) are characterized by the disregard of information and therefore they are cognitively simpler to perform than BUS (1). Strategies, which combine information additively are discussed in several contexts, such as joint probabilities (Sundh, 2019), conjunctive probabilities (Juslin, Lindskog, & Mayerhofer, 2015), and Bayesian reasoning (Cohen & Staub, 2015; Juslin et al., 2009; Lopes, 1985; Shanteau, 1975). Additive strategies combine all probability information and can approximate multiplicative strategies in some situations. These strategies assume that the individual determine posteriors by a complex weighted sum of all probabilities (e.g., Cohen & Staub, 2015; Juslin et al., 2009):

$$P(H_i|E) = \alpha \cdot P(H_i) + \beta \cdot P(E|H_i) + \gamma \cdot P(E|\neg H_i) \qquad (4)$$

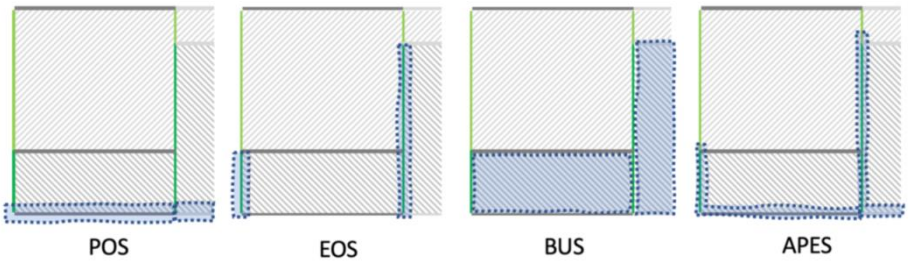
It appears rather implausible that a teacher processes such a complex set of information on probabilities and regression weights intuitively when judging the probability of misconceptions based on their students' responses. Therefore, we assume a more cognitively plausible additive strategy (see also Shanteau, 1975):

$$P(H_i|E) \propto \tfrac{1}{2}\big(P(H_i) + P(E|H_i)\big) \qquad \text{(Averaging-Prior-Evidence, APES)} \quad (5)$$

DOI: https://doi.org/10.23770/

Mathematically, APES can be regarded as an approximation, since it qualitatively reflects the "magnitude" of the product in (1). This can be illustrated within a rectangular visualization of the Bayesian situation (Figure 3). Usually this visualization is used to support normative Bayesian reasoning BUS (e.g., Böcherer-Linder & Eichler, 2017). However, it proves also useful to illustrate the (mathematically incorrect) strategies: While POS remains on the prior probabilities (gray horizontal bars) and EOS only uses the likelihoods (green vertical bars), BUS correctly regards the interaction of prior probabilities and likelihoods (multiplication). APES also considers both, but approximates the interaction by averaging the probabilities.

Fig. 3: The highlighted areas show how the strategies POS, EOS, BUS, and APES take into account the probability information – priors (gray horizontal bars) and likelihoods (green vertical bars).



Cognitively, it is simpler to derive an additive average of two magnitudes (e.g., ½ (90% + 20%)) than a multiplicative interaction and therefore this strategy is cognitively more plausible. This is expressed in the intuitive mental model for averaging subjective probabilities (Khemlani et al., 2015) and amounts to a "take-the-middle heuristic".

Empirically, one can find indicators for such additive strategies in the literature on Bayesian reasoning: For instance, the responses of the participants in the study by Cohen and Staub (2015) could be better modeled as an additive combination of multiple probabilities than as a multiplicative combination of probabilities. Shanteau (1975) showed that when updating probability estimations based on (non-informative) evidence, the probability updates of the participants suggest averaging (APES) instead of multiplying (BUS).

3 Research question

How can exact Bayesian reasoning be approximated when representations of probabilities are non-numerical and fuzzy, and when information processing relies on cognitively feasible heuristics? We developed our argument in the context of teachers' diagnostic judgments on students' misconceptions. However, our approach can also be seen from a more general point of view of judging under uncertainty.

Building on cognitive plausibility and empirical evidence, we argue that for teachers' diagnostic judgments, one may assume a strategy, which averages probabilities of priors and evidence (APES). This strategy approximates exact Bayesian reasoning by reducing

DOI: https://doi.org/10.23770/

complexity without neglecting information. In two computational explorations, we study the potential effectiveness of APES by investigating the following research question:

For which types of situations (i.e., constellations of prior and likelihood values) can the averaging-priors-and-evidence strategy (APES) be regarded as an adequate and effective approximation of the Bayesian update strategy (BUS) and as a substantial improvement with respect to the information-neglecting strategies "prior only" and "evidence only" (POS, EOS)?

Since we embrace a perspective of ecological rationality, we discuss the adequacy of the approximative strategy against the background of teachers' diagnostic judgments and thus focus on situations of judging a certain piece of evidence (a student's response), which is indicative of one of several possible hypotheses (students' misconceptions) (for an empirical investigation of this situation, cf. Leuders & Loibl, 2020).

4 Methods and Results

4.1 Study 1: Inferring from evidence whether a student does or does not have a misconception

4.1.1 Methods of Study 1

In the scenario for Study 1, we assume a situation with the two hypotheses H_1, H_2 that a student does or does not have a misconception and evidence that is indicative of the misconception (high likelihood for H_2) and reduces plausibility for the hypothesis that the student has no misconception (low likelihood for H_1). Accordingly, we explore constellations with the following set of values:

$$P(H_1), P(H_2) \in [0;1], \sum P(H_i) = 1$$

$$P(E|H_1) \in [0.10; 0.40], P(E|H_2) \in [0.60; 0.90]$$

For the computational simulation of the various strategies, we assume the following cognitive process: the probability information (priors $P(H_i)$ and likelihoods $P(E|H_i)$), and the evidence *E* are available. The goal is to decide which of the two posterior hypotheses $P(H_i|E)$ has the higher probability: a student does or does not have a misconception. To that purpose one of the following strategies is activated (the $\gtrless$-sign meaning "compare and decide for the larger"):

$P(H_1) \gtrless P(H_2)$ (POS)

$P(E|H_1) \gtrless P(E|H_2)$ (EOS)

$P(H_1)P(E|H_1) \gtrless P(H_2)P(E|H_2)$ (BUS)

DOI: https://doi.org/10.23770/

$$\frac{1}{2}[P(H_1) + P(E|H_1)] \gtrless \frac{1}{2}[P(H_2) + P(E|H_2)] \qquad \text{(APES)}$$

A normalization factor (e.g., $\Sigma_i P(H_i)P(E|H_i)$ for BUS), which is necessary to attain a value of the posterior probability, is not required for a decision between the hypotheses, since it is identical for both sides of the comparison in each strategy.

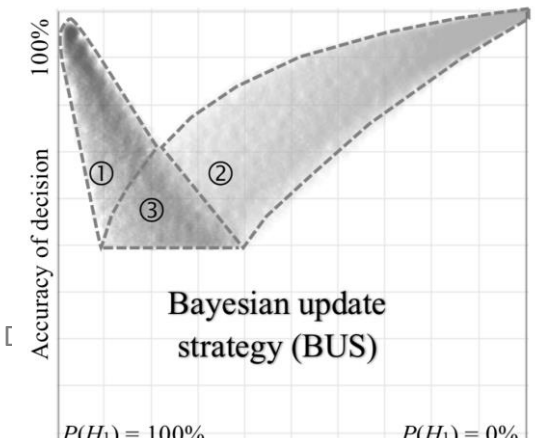
The simulation and the graphical representation of the results was implemented in Cinderella (Richter-Gebert & Kortenkamp, 2011), a programming environment for numerical calculation and visualization (source code available from the authors).

4.1.2 Results of Study 1

To evaluate and compare the various investigated reasoning strategies, we graphically display the outcome (the decision for a hypothesis) throughout the whole parameter space in a way that makes the phenomena most salient (see Figure 4-6): Each dot represents a decision based on the respective strategy for a certain set of parameters in the probability space. The decision depends on the comparison of the posterior probabilities for H_1 ("the student has no misconception") and H_2 ("the student has a misconception"). The values of the prior probabilities are represented by the *x*-coordinate. The accuracy of the decision is the corresponding posterior and is represented by the *y*-coordinate. For each value on the *x*-coordinate there is an interval of values on the *y*-coordinate due to the variation of the likelihood values. Numbers ①-③ indicate regions of decisions as explained in the text.

Figure 4 presents the results of the optimal Bayesian decision (BUS) when evidence is indicative for H_2, that is the response is erroneous. In region ①, BUS leads to a decision for H_1because the prior probability for H_1 is high. That is, despite the erroneous response, the teacher should assume that the error was not caused by a misconception. When the prior probability for H_2 is middle to high as in region ②, BUS results in a decision for H_2 due to the evidence for H_2. That is, the teacher should assume that the student has a misconception and consequently should intervene. For a certain prior probability constellation (region ③) BUS leads to either H_1 or H_2 depending on the likelihood values. In this region, it seems most important for a teacher to gather more information (e.g., posing more diagnostic tasks).

Fig. 4: Accuracy of the decision based on the BUS strategy.



BUS is the mathematical optimal update strategy and can be used for decisions by selecting the hypothesis with the highest posterior probability (in case of two hypotheses, the hypothesis with a probability ≥50%). The decision accuracy of the other strategies (EOS, POS, APES) depends on the accuracy of BUS and can never exceed the value for BUS. If a decision coincides with BUS, the accuracy of the decision is the same as for BUS. If the decision deviates from BUS, the accuracy of the decision corresponds to the lower decision accuracy for the opposite hypothesis following BUS (not displayed in Figure 4). For example, the strategy EOS always results in a decision for H_2 ("the student has a misconception"): A teacher following this strategy only considers the evidence, which in our case is an erroneous response. For certain prior and likelihood values (④ in Figure 5), this decision coincides with the optimal decision according to BUS with the accuracy $P_{\mathrm{BUS}}(H_2|E) \geq 50\%$. For other prior and likelihood values, H_1 has the higher probability according to BUS. Here, EOS deviates from BUS (⑤ in Figure 5) and, thus, the decision based on EOS has the (lower) accuracy $P_{\mathrm{BUS}}(H_2|E) \leq 50\%$.

Figure 5 presents the comparison of the decisions based on the information neglecting strategies (POS, EOS) and the optimal Bayesian decision (BUS) when evidence is indicative of H_2, i.e., the response is erroneous. The decision of the information neglecting strategies coincides with Bayesian reasoning for certain constellations of priors and likelihoods in regions ④. The decision deviates in the broad regions ⑤ of prior values due to the disregard of evidence (POS) or priors (EOS), resulting in low decision accuracy (i.e., posterior probability according to BUS below 50%) in these regions.

Fig. 5: Accuracy of the decision based on information neglecting strategies POS (green, l.h.s.) and EOS (red, r.h.s.) in comparison to the accuracy of BUS (grey in the background).

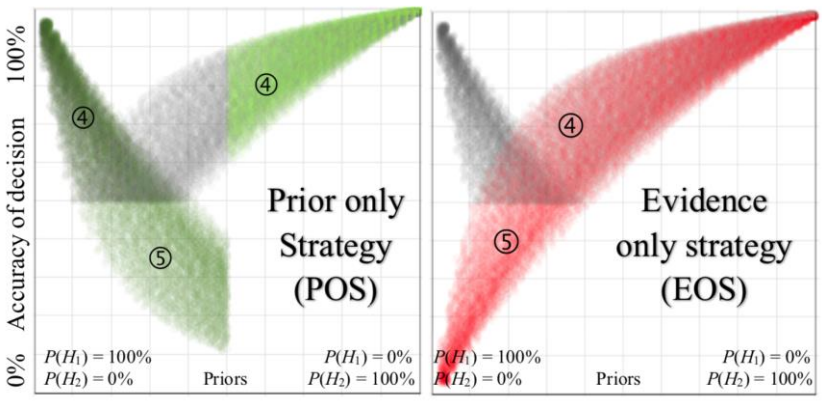
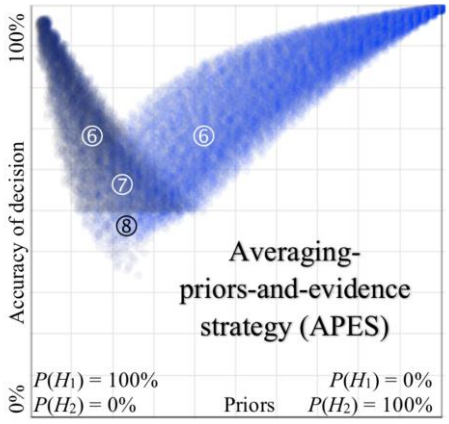


Figure 6 presents the comparison of the decision based on the averaging-priors-and-evidence strategy (APES) and the optimal Bayesian decision (BUS) when evidence is indicative for H_2, i.e., the response is erroneous. The decision based on APES coincides with Bayesian reasoning for most constellations of priors and likelihoods as shown by the broad regions ⑥. In the prior region where the BUS decision for H_1 or H_2 depends on the likelihood constellations (compare region ③ in Figure 4), APES coincides with BUS in region ⑦ and deviates from BUS in region ⑧, depending on the likelihood constellations. Moreover, the deviations only result in small reductions of the accuracy in comparison to the accuracy of the Bayesian decision. The small magnitude of the accuracy reduction is due to the fact that the region with deviations falls in the region with the smallest accuracy of the Bayesian decision, close to 50%. As argued above, in this region a teacher would anyway gather more information before rendering a judgment.

Fig. 6: Accuracy of the decision based on the averaging-priors-and-evidence strategy (APES, blue) in comparison to the accuracy of the decision based on BUS (grey in the background).



4.2 Study 2: Inferring multiple causes from evidence

4.2.1 Methods of Study 2

In the context of teacher judgment, the situation is often more complicated than deciding between two hypotheses. As an extension to the scenario of Study 1, we explored the more complex situation of three hypotheses (e.g., three different misconceptions). In the context of decimals, the three most prevalent misconceptions are the whole-number misconception, the ignore-decimal-point misconception, and the

DOI: https://doi.org/10.23770/

shorter-is-larger misconception. While students with the whole-number misconception and with the ignore-decimal points usually respond that 4.8 is smaller than 4.63 (because 8 is smaller than 63 or 48 is smaller than 463), students with the shorter-is-larger misconception usually answer that 4.8 is larger than 4.63 as 4.8 only has one digit and 4.63 has two digits (cf. Moloney & Stacey, 1997). Thus, in this example a correct response is sensitive for the shorter-is-larger misconception but is ambiguous with regard to the other hypotheses. Other task-response constellations are sensitive for the whole-number misconception or the ignore-decimal-point misconception (cf. Leuders & Loibl, 2020 for example tasks). That is, in Study 2 we explore a situation where the evidence is sensitive and specific for only one of the hypotheses (i.e., high likelihood only for one of the hypotheses, cf. Figure 7, l.h.s.). We therefore explore situations with the following values:

$$P(H_1), P(H_2), P(H_3) \in [0;1], \sum P(H_i) = 1$$

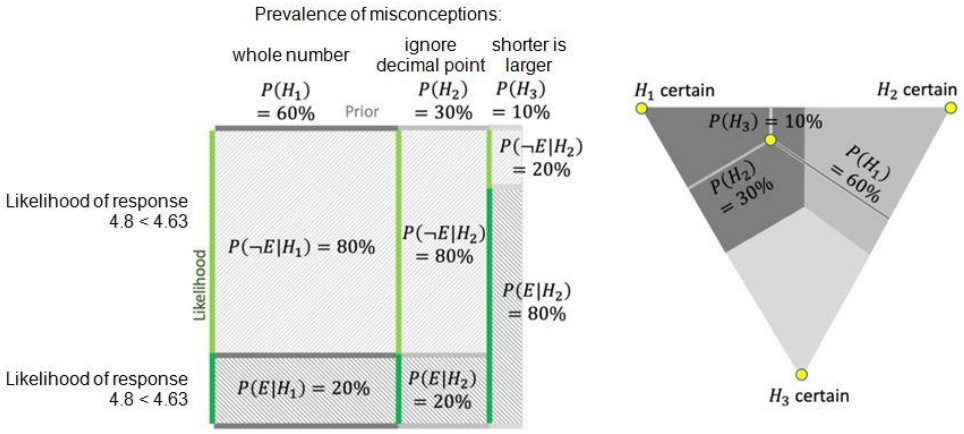
$$P(E|H_1), P(E|H_2) \in [0.10; 0.40],$$

$$P(E|H_3) \in [0.60; 0.90]$$

The simulation algorithm of the strategies corresponds to Study 1: A decision for one of the three hypotheses is reached by deciding in favor of the highest posterior probability according to each strategy. The representation of the results, however, has to be adapted to the affordances of the higher dimensionality of three hypotheses. To that purpose, we use barycentric homogeneous coordinates. The diagram, which we call “hypothegon” (cf. Figure 7, r.h.s.; de Finetti, 2017; Jøsang, 2016; Leuders & Loibl, 2020), extends the hypothesis line [0,1] for two hypotheses to a triangular prior space for three hypotheses.

Fig. 7: A situation with three hypotheses and their likelihoods (low likelihood 20% for H_1 and H_2, high likelihood 80% for H_3) with respect to evidence *E* (l.h.s). Any set of prior probabilities (e.g., 60%, 30%, 10%) can be regarded as convex coordinates for a unique locus within a triangle (“hypothegon”) (r.h.s.).

DOI: https://doi.org/10.23770/



4.2.2 Results of Study 2

As outcome of the various strategies, we again display the accuracy of each decision, depending on the value of the prior probabilities. In order to attain a more pronounced geometry, we chose to only display the results for a fixed likelihood (80%), keeping in mind that the likelihood interval leads to an interval of the accuracy value for fixed prior probabilities.

Figure 8 gives a comprehensive picture of the averaging-prior-and-evidence strategy (APES, blue) compared to the Bayesian strategy (BUS, grey) when evidence is indicative of H_3 as in the example provided above: correct answer 4.8 > 4.63 and the three hypotheses whole-number misconception, ignore-decimal-point misconception, and shorter-is-larger misconception. In most regions, both strategies APES and BUS *coincide* with high decision accuracy in regions ① with a large prior probability for one of the hypotheses (H_1, H_2, or H_3) and medium decision accuracy for region ② with similar prior probabilities for all hypotheses ($H_1 \approx H_2 \approx H_3 \approx 30\%$). The only prior region in which APES *deviates* from BUS, is region ③ with low prior probability for H_3 and similar prior probabilities for H_1 and H_2 ($H_1 \approx H_2 \approx 50\%$). In our setting, the evidence is contra-indicative for H_1 and H_2 (e.g., whole-number misconception and ignore-decimal-point misconception) with ambiguous likelihoods for these hypotheses (e.g., correct answer 4.8 > 4.63). This ambiguity comes into effect only in regions with similar and rather high prior probabilities. Due to the ambiguity the accuracy of APES drops from 70%-50% to 50%-0% in region ③. Further variation of the likelihood values (not displayed here) does not alter the general picture but only slightly increases the region ③ in which APES deviates from BUS.

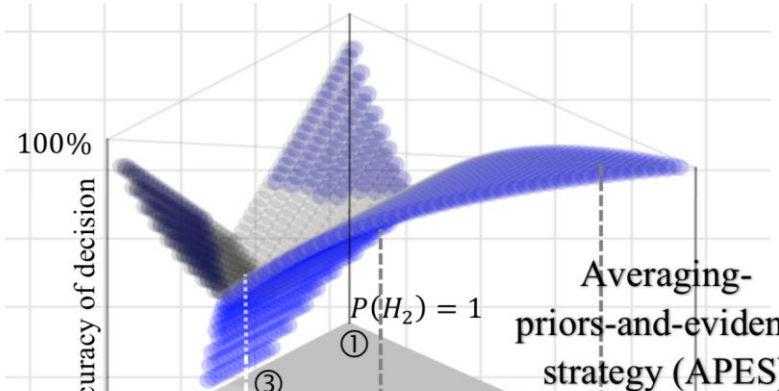
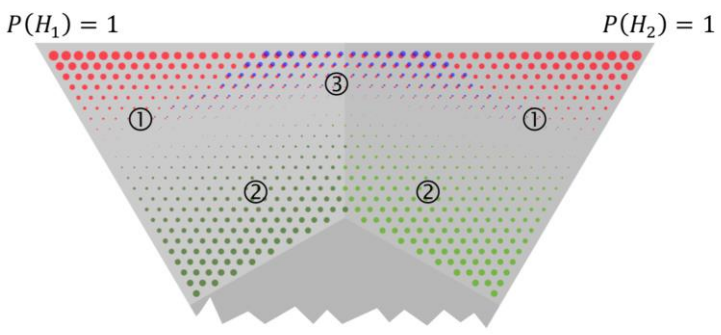


Fig. 8: Accuracy of the decision based on the averaging-priors-and-evidence strategy (APES, blue) in comparison to the accuracy of BUS (gray) with regard to three hypotheses. The prior probabilities are coordinates within the triangle (cf. Figure 7).

In order to compare the deviations of the decision based on the heuristic strategies EOS, POS, and APES from the ideal Bayesian decision (BUS), we calculate the differences in decision accuracy (in %) and display the resulting "error distribution" over the whole prior space (hypothegon) for fixed likelihoods (20%, 20%, 80%) (Figure 9). For the information neglecting strategies EOS and POS, there are – similar to Study 1 with two hypotheses (Figure 5) – large deviations for EOS (displayed in red) in region ① with extreme priors contrary to the evidence (EOS decision H_3, BUS decision H_1 or H_2) and large deviations for POS (displayed in green) in region ② with priors slightly contrary to the evidence (POS decision H_1 or H_2, BUS decision H_3).

In contrast, for APES (displayed in blue) the deviations are much smaller and restricted to region ③ with extreme priors contrary to the evidence and ambiguous likelihoods for H_1 and H_2. For more extreme likelihood parameters (e.g., 10%, 10%, 90%) this region is closer to the extreme boundary and for less extreme likelihood parameters (e.g., 40%, 40%, 60%) it approaches the middle but with only very small values of deviation (<10%, not displayed here). As shown in Figure 9, there are no deviations for any strategy, when priors and evidence suggest identical decisions.

Fig. 9: Deviations of decision accuracy for EOS, POS, and APES from the ideal Bayesian decision. The size of the dots corresponds to the magnitude of the deviations with the largest dots (e.g., in the corners) corresponding to a 100% deviation.



5 Discussion

5.1 Findings and limitation of the studies

In our study, we investigated heuristic strategies in Bayesian reasoning with respect to their accuracy, depending on the set of values of prior probabilities of the hypotheses and likelihoods of the evidence. In addition to the mathematically optimal Bayesian strategy (BUS) and the well-documented heuristic strategies, that neglect either evidence or prior information (EOS, POS), we proposed an averaging strategy (APES) and underpinned its plausibility based on cognitive and empirical arguments from literature (e.g., Cohen & Staub, 2015; Khemlani et al., 2015; Shanteau, 1975). The exploration of a broad parameter space (two and three hypotheses with priors from [0;1], evidence with likelihoods from [0.6;0.9]) yielded the following insights:

- Processing of evidence only (EOS, prior neglect) or priors only (POS) results in low accuracy compared to the Bayesian update strategy (BUS) in broad regions of prior probabilities. These deviations occur when priors and evidence suggest divergent decisions (⑤ in Figure 5; ① & ② in Figure 9).

- Averaging prior and evidence probabilities (APES) is a good approximation for Bayesian reasoning, leading to identical decisions for most values of priors and likelihood (⑥, ⑦ in Figure 6; ① & ② in Figures 8 & 9).

- Deviations of APES from BUS remain small and occur in confined regions of prior values. We found two situations (ranges of values) where the decision based on APES deviates from the BUS decision: (1) similar values of competing posteriors (⑧ in Figure 6), which is also the most inconclusive (i.e., least accurate) situation in exact Bayesian reasoning, (2) ambiguous hypotheses (here: hypotheses with similar low likelihoods) sharing high prior probabilities (③ in Figures 8 & 9).

5.2 Interpretation of the strategies in the context of teachers' diagnostic judgments

From a perspective of ecological rationality, heuristic strategies possibly constitute viable and efficient cognitive strategies in certain situations (cf. Simon, 1955; Gigerenzer

DOI: https://doi.org/10.23770/

& Goldstein, 1996). To investigate the heuristic strategies with regard to their ecological rationality, we specify the situational framing as teachers' diagnostic judgments on students' potential misconceptions (with certain prior probabilities) based on responses to tasks (evidence). We assume that teachers usually do not have the precise probability values available, but only noisy estimates, such as a misconception is very frequent/frequent/rare etc. or students with the misconception usually/sometimes respond in this way. The computational model here does not explicitly reflect the noisiness of analog non-numerical mental representations of subjective probabilities (cf. research on magnitude and analogue representations: Gallistel, 2011; Khemlani et al., 2015; Leibovich, Katzin, Harel, & Henik, 2017 or models for uncertainty in probability estimation: Jøsang, 2016) as modeling the noisiness of the probability estimates would blur the figures. Therefore, when interpreting the results, we have to account for the noisiness of the probability estimates by interpreting the numerical values as centers of approximative intervals, e.g., 25% represents ca. 20-30%.

Within the situational framing, a teacher may judge for instance, what is the plausibility of the hypothesis H_2 that a student has a whole number misconception after a response to a task (e.g., "4.8 < 4.63")? A teacher ideally considers all the information (as subjective probabilities not necessarily represented numerically) and processes this information to arrive at a decision on the misconception with the highest posterior probability. Given the complexity of the situation, it is likely that a teacher applies one of the heuristic strategies. For the case of teachers' decisions, we put forward the following interpretations:

- As teachers use diagnostic tasks in order to receive evidence about their students' skills and misconceptions, teachers are not likely to ignore this evidence. We therefore argue that teachers do not apply a priors-only strategy (POS) when evidence is available.

- When teachers apply an evidence-only strategy (EOS), they focus on students' errors and tend to react immediately with instruction (Herppich, Wittwer, Nückles, & Renkl, 2016; Phelps-Gregory & Spitzer, 2018). This may be an inaccurate diagnostic decision for misconceptions with low base rate (and inefficient considering restricted instructional time). However, when applied as screening for further diagnostic interaction, the strategy may be appropriate.

- While Bayesian reasoning is computationally rather complex, a simple averaging strategy (APES) seems not only cognitively plausible, but also appropriate and feasible in situations with quick on-the-fly assessment. Our results indicate that they lead to optimal (Bayesian) decisions in most cases, and only deviate in situations of ambiguity (similar posteriors). In these cases, an appropriate decision for teachers would be to *not* decide on one misconception or the other but to resume assessment. That is, teachers should select further diagnostic tasks that help to distinguish between the hypotheses that could not be distinguished by the first task. In our example, the wrong answer 4.8 < 4.63 is ambiguous with regard to the whole number misconception and the ignore decimal misconception. These

DOI: https://doi.org/10.23770/

misconceptions can be distinguished by the task compare 3.7 and 3.02, which will probably be answered correctly given the whole number misconception (7 > 2), but wrongly given the ignore decimal points (37 < 302).

Our analysis gives strong support to consider averaging and considering all information (APES) as a promising heuristic strategy for Bayesian decision situations in general, and ecologically valid for teachers' diagnostic judgments. In order to further support the plausibility of APES, we are working towards modeling the cognitive representation of non-numerical probabilities (cf. Khemlani et al., 2015) and cognitive processing according to APES.

In our analysis, we chose a situation that was rather clear-cut with respect to the theory on student thinking and the relation of latent constructs and manifest errors. We posit that such situations can be found throughout many topics in different domains. However, this does not imply that such situations are generic or even typical of diagnostic judgments. In some domains, the relation of student thinking and student behavior may be less direct, especially when categories like "correct" or "incorrect" are not applicable, and a teacher has to judge the students' individual interpretations of texts or situations (e.g., language teaching or literature). Here, a model with a set of disjunctive hypotheses as in Bayesian reasoning would not be considered an adequate cognitive model for diagnostic judgments. One would rather use approaches such as mental models of social situations (Johnson-Laird, 1983) or perspective taking (Nickerson, 1999).

5.2 Further research

In spite of the ecological rationality of APES, we do not posit that our study implies actual prevalence of APES in the context of teacher diagnostic judgments: We have no information on teachers' application of such a strategy during teaching. Furthermore, teachers' diagnostic cognitive processes are far more complex than the focus of our analysis (e.g., single- vs. multiple-cues judgments) and require more comprehensive models (Herppich et al., 2018; Loibl et al., 2020). We therefore consider the most pressing aim for further research to ascertain empirically the validity of the averaging strategy (APES) in non-numerical situations, since empirical support is still rare. Deriving empirical evidence for APES in non-numerical situations may be challenging due to the small deviations of APES from BUS. A promising route to this goal is to extend the model from deciding on the largest posterior to estimating posterior values. Following this route, we gained first experiences with intervention studies in which we systematically prompted the processing of all probability information similar to APES in the context of teacher judgments, which however did not focus on detecting APES, but other strategies (Leuders & Loibl, 2020). Based on these experiences and given the complexity of judgments in the classroom, investigating the prevalence of APES only seems realistic in laboratory studies that allow controlling confounding variables and selecting probabilities that actually lead to different judgments, when applying APES or BUS.

DOI: https://doi.org/10.23770/

In addition, it would also be interesting to further investigate the heuristic strategies within more complex situations (multiple cues, further sets of values for sensitivity of evidence), and to better understand the role of ambiguity. Finally, the relevance of the cognitive approach for practical judgment situations, can be better discussed, when we have findings from research designs that actually are framed or even embedded in (real or artificial) classroom situations.

Acknowledgments

This work was funded by the Ministry of Science, Research and Arts of Baden-Wuerttemberg within the Research Training Group “Diagnostic Competences of Teachers” (DiaKom).

References

Böcherer-Linder, K., & Eichler, A. (2017). The impact of visualizing nested sets. An empirical study on tree diagrams and unit squares. *Frontiers in psychology 7*, 241.

Cohen, A. L., & Staub, A. (2015). Within-subject consistency and between-subject variability in Bayesian reasoning strategies. *Cognitive Psychology, 81*, 26–47.

De Finetti, B. (2017). *Theory of Probability: A Critical Introductory Treatment*. New York, NY: John Wiley & Sons.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal Representation of Human Judgment* (pp. 17-52). New York, NY: Wiley.

Gallistel, C.R. (2011). Mental magnitudes. In S. Dehaene, E.M. Brannon (Eds.), *Space, time and number in the brain: Searching for the foundations of mathematical thought* (pp. 3-12). London: Elsevier.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review, 103*(4), 650.

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

Heitzman, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., & Fischer, F. (2019). Facilitating Diagnostic Competences in Simulations: A Conceptual Framework and a Research Agenda for Medical and Teacher Education. *Frontline Learning Research*, *7*(4), 1–24.

Herppich, S., Praetorius, K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., Behrmann, L., Böhmer, M., Ufer, S., Klug, J., Hetmanek, A., Ohle, A., Böhmer, I., Karing, C., Kaiser, J., &

DOI: https://doi.org/10.23770/

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2016). Expertise amiss: Interactivity fosters learning but expert tutors are less interactive than novice tutors. *Instructional Science, 44,* 205–219.

Hoppe, T., Renkl, A., & Rieß, W. (2020). Förderung von unterrichtsbegleitendem Diagnostizieren von Schülervorstellungen durch Video und Textvignetten. *Unterrichtswissenschaft, 48*, 573–597.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* . Harvard University Press.

Jøsang, A. (2016). Generalising Bayes' theorem in subjective logic. *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)* (pp.462–469). Baden-Baden.

Juslin, P., Lindskog, M., & Mayerhofer, B. (2015). Is there some-thing special with probabilities? Insight vs. computational ability in multiple risk combination. *Cognition, 136*, 282–303.

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review, 116*(4), 856–874.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*(3), 582–591.

Khemlani, S. S., Lotstein, M., & Johnson-Laird, P. N. (2015). Naive probability: Model-based estimates of unique events. *Cognitive Science, 39*(6), 1216–1258.

Krolak-Schwerdt, S., Pit-ten Cate, I. M., & Hörstermann, T. (2018). Teachers' judgments and decision-making: Studies concerning the transition from primary to secondary education and their implications for teacher education. In *Assessment of Learning Outcomes in Higher Education* (pp. 73–101). Springer, Cham.

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From 'sense of number' to 'sense of magnitude' – the role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences, 40*, E164.

Leuders, T., & Loibl, K. (2020). Processing probability information in non-numerical settings – teachers' bayesian and non-bayesian strategies during diagnostic judgment. *Frontiers in Psychology, 11,* 678.

Loibl, K., & Leuders, L. (2020). "Take the middle" – Averaging prior and evidence as effective heuristic in bayesian reasoning. In S. Denison., M. Mack, Y. Xu, & B.C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 1764–1770). Cognitive Science Society.

Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teaching and Teacher Education, 91,* 103059.

Lopes, L. L. (1985). Averaging rules and adjustment processes in Bayesian inference. *Bulletin of the Psychonomic Society, 23*, 509–512.

DOI: https://doi.org/10.23770/

Mandel, D. R. (2014). The psychology of bayesian reasoning. *Frontiers in Psychology, 5*, 1144.

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. R. (2003). Naive and yet enlightened: From natural frequencies to fast and frugal decision trees. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspective on reasoning, judgment, and decision making* (pp. 189–211). Chichester, UK: Wiley.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. Psychological Bulletin, 125, 737–759.

Phelps-Gregory, C. M., & Spitzer, S. M. (2018). Developing prospective teachers' ability by classroom intervention: Replicating a classroom intervention. In T. Leuders, J. Leuders, & K. Philipp (Eds.), *Diagnostic Competence of Mathematics Teachers: Unpacking a complex construct in teacher education and teacher practice* (pp. 223–240). New York: Springer.

Richter-Gebert, J., & Kortenkamp, U. H. (2000). *User manual for the interactive geometry software cinderella*. Springer Science & Business Media.

Shanteau, J. (1975). Averaging versus multiplying combination rules of inference judgement. *Acta Psychologica, 39*, 83–89.

Simon, H. A. (1995). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118.

Südkamp, A. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education, 76*, 181–193.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of educational psychology, 104*(3), 743.

Sundh, J. (2019). The Cognitive Basis of Joint Probability Judgments. Processes, Ecology, and Adaption. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences, 166*. Uppsala: Acta Universitatis Upsaliensis.

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: an account of deviations from Bayes' theorem and the additivity principle. *Memory & Cognition, 30*(2), 171–178.

Weber, P., Binder, K., & Krauss, S. (2018). Why can only 24% solve Bayesian reasoning problems in natural frequencies: Frequency phobia in spite of probability blindness. *Frontiers in Psychology, 9*, 1833.

Witzigmann, St., Sachse, St. (2021). Diagnostic competencies of prospective teachers of French as a foreign language: judgement of oral language samples. *RISTAL, 4*, 70–86.

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition, 98,* 287–308.

Katharina Loibl

DOI: https://doi.org/10.23770/

is a professor for interdisciplinary research on learning and instruction at the University of Education Freiburg, Germany with a research focus on learning mechanisms and instructional designs. She is co-speaker of the research training group “DiaKom” within which this research was conducted.

Timo Leuders

Timo Leuders is a professor for mathematics education at the University of Education in Freiburg, Germany with a research focus on teaching and learning in secondary education and teacher professionality. He is co-speaker of the research training group “DiaKom” within which this research was conducted.

DOI: https://doi.org/10.23770/