PROTO-BAYESIAN REASONING OF YOUNG STUDENTS IN THE CONTEXT OF DRAWING WITH REPLACEMENT

Torsten Fritzlar[1], Denise Resche[2]

[1]Martin Luther University Halle-Wittenberg, [2]Leipzig University

Bayesian reasoning is a core element of statistical literacy but remains difficult even for adults. This qualitative exploratory study examines proto-Bayesian reasoning of 9–12-year-old students in a simple Bayesian model selection task with two bags and sequential draws of cubes with replacement. Using clinical interviews and probability scales, third- to fifth-grade students estimated and updated likelihoods after 5, 10, and 15 draws. Results show that many students intuitively adjust likelihoods in response to new information. The findings indicate that successful proto-Bayesian reasoning is possible in the age group studied and probability calculation skills are not necessary for proto-Bayesian reasoning.

INTRODUCTION

"Most important decisions men make are governed by beliefs concerning the likelihood of unique events" (Tversky & Kahneman, 1973, p. 231). This statement, now more than 50 years old, is even more relevant in our increasingly data-driven society, where statistical and probabilistic information is an important basis not only for political and professional decision-making, but also for individual decision-making. Therefore, understanding, critically evaluating, and deciding on probabilistic information are essential components of statistical literacy (e.g., Gal & Geiger, 2022). A reasonable handling of conditional probabilities is an important prerequisite for coping with risk and uncertainty. Especially Bayesian reasoning, which is the process of updating likelihoods for hypotheses based on new information as a source for conclusions and decisions, is crucial for many professions and also in everyday life.

While it was initially assumed that the theory of probability is "only common sense reduced to calculus" (Laplace, 1902, p. 196), later studies have shown that many adults fail to reason according to the laws of probability, even if they are well educated or professional experts. Bayesian reasoning in particular has often proven to be especially challenging (e.g., Gigerenzer et al., 2021). The importance of Bayesian reasoning on the one hand and its particular difficulty on the other leads to important tasks for research in mathematics education. Among other things, it seems important to examine students' proto-Bayesian reasoning in order to lay the foundations for developing effective learning opportunities. Due to the particular difficulty of proto-Bayesian reasoning, for a very long time there was no research on this topic involving primary school students. Even today, there is still no experience with so-called model selection tasks. The study presented in this paper aims to contribute to closing this gap by

2026. In H. Krzywacki, M. Hähkiöniemi, E. Haataja, M.S. Hannula & P. Hästö (Eds.). *Proceedings of the 49th Conference of the International Group for the Psychology of Mathematics Education: Research Reports*, pp. 1219-1226. PME.

investigating how young students estimate the likelihood from which of two bags colored cubes were drawn based on 5, 10 and 15 draws with replacement.

(PROTO-) BAYESIAN REASONING

Bayes' rule, initially largely ignored and later rediscovered by Laplace, has become a fundamental idea in probability and the basis of several statistical domains. The importance of the Bayesian approach lies in the fact that it incorporates prior knowledge into the analyses and allows an update of the probability on the basis of new data, a learning from experience. Therefore, and because it assigns probabilities to hypotheses instead of data, the Bayesian approach is seen as more in line with our natural mode of reasoning (Fornacon-Wood et al., 2022). Because of this significance, Bayesian reasoning is very interesting for psychological and mathematics education research. This is especially true for proto-Bayesian reasoning, which refers to the intuitive and often informal ways in which subjects who have had no instruction in Bayes' rule arrive at their conclusions in Bayesian situations.

Research in psychology has repeatedly found over several decades that subjects, even academically educated adults, struggle when dealing with Bayesian situations. This applies to both types of Bayesian tasks that are distinguished in the research literature: parameter estimation and model selection (Gallistel et al., 2014). The first involves judgements of parameters of a probability distribution, the second requires the subject to choose between a small number of distributions or to compare their likelihoods given some observed outcomes.

Due to the empirically proven difficulties of adults, by the end of the 20th century, no single study had investigated whether children are able to make judgments according to Bayes' rule (Gigerenzer et al., 2021). For the first type of task, however, it has been shown in recent years that difficulties can be significantly reduced even among primary school students when working with natural frequencies and iconic representations (e.g., Gigerenzer & Hoffrage, 1995; Gigerenzer et al., 2021). However, as far as we know, there is still no empirical research on proto-Bayesian reasoning of young students for the second type of tasks.

RESEARCH INTEREST

The study presented in this paper addresses the described desideratum in the context of a Bayesian situation concerning model selection that is as simple as possible. It examines how young students estimate the probability from which of two bags colored cubes were drawn based on the known fillings of the bags and 5, 10, and 15 given draws with replacement (cf. Fig. 1). Even though these estimates can be interpreted as likelihood judgments in the Bayesian sense, the more general term "probability" will be used in the context of the empirical study. Since the transition from primary to secondary school (9-12-year-old students) is characterized by dynamic cognitive development in students, it also seems interesting to examine whether, even without

addressing the topic in mathematics classes, there are indications of development in proto-Bayesian reasoning across different grades in this age range (3rd-5th grade).

Therefore, we are interested in the following research questions concerning the described Bayesian situation:

1) To what extent do students arrive at probability estimates that are appropriate to the information currently available?
2) How do the students justify their probability estimates?
3) What similarities and differences are there between students' solutions in the third, fourth, and fifth grades?

Due to space constraints, we cannot address the second question in this paper.

METHODICAL ASPECTS

Data Collection

Since there is little empirical evidence available to date, it seemed appropriate to conduct a qualitative exploratory study. Three schools (two in a large city and one in a rural area in Germany) were approached to recruit participants. A total of 11 third graders, 21 fourth graders, and 8 fifth graders were recruited. There were no further criteria for composing the study group. Although subjects were selected based on their interest in participating, the participants were not exclusively high-achieving students or those with a keen interest in mathematics.

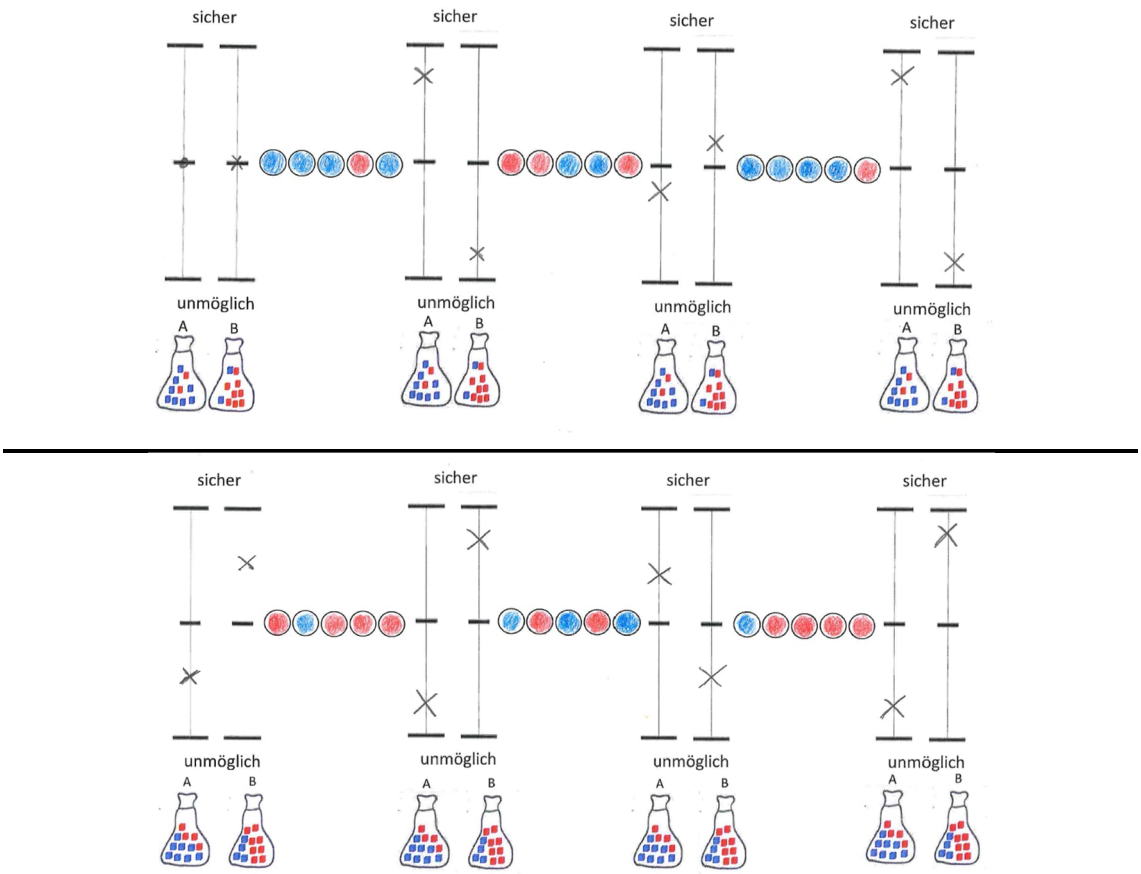
In semi-standardized clinical interviews (e.g., Hunting, 1997), students were given three video-based tasks. In this paper, we can present only two of them, which are structured analogously: Two bags, A and B, are filled with red and blue cubes. The cubes are counted aloud in front of the students in the video, and the contents of the bags are also shown on the respective worksheet. Hidden from the students, one of the two bags is selected "at random," from which 15 cubes are drawn with replacement one after the other. Before the first draw and after every five draws, the students are asked to estimate the probability that the cubes will be drawn from bag A or bag B.

In Germany, national standards for primary school mathematics education do not require the use of numerical probability values. However, all students should be able to estimate probabilities and compare them with each other (KMK, 2022). In this study, probability estimates are therefore to be given using continuous probability scales from "certain" to "impossible." Such scales have low language requirements and support the quantification of probability without requiring calculations. The students' worksheets show a scale for each bag and each estimate. The scales are arranged vertically to avoid left-right confusion and to illustrate increases and decreases in probability in a manner analogous to a thermometer (cf. Fig. 1).

Figure 1 shows, as an example, how fourth grader Maurice completed the two tasks. Like all students, he recorded the draws of the blue and red cubes presented in the video

(by coloring the corresponding fields) and indicated his probability estimates by marking them with crosses on the scales. (The estimates “certain” and “impossible” should be marked in a special way.) The two worksheets also make it clear that in the second task, the bags were filled with a different ratio of colors (8b+2r & 2b+8r in task 1; 8b+4r & 4b+8r in task 2), but in the drawings of cubes, only the role of the colors was swapped (1 of 5, 3 of 10, 4 of 15 in both tasks). Based on this, students could also establish relationships between the two tasks and use them for their probability estimates.

Fig. 1: Results of Maurice (4th grade; “sicher” means certain, “unmöglich” means “impossible”)



On the one hand, the Bayesian situation used in the study is particularly simple, as only two bags (or source distributions in a more abstract sense) are possible. However, this also means that the mathematically correct likelihoods add up to 1 and the corresponding markings on the probability scales are symmetrical to each other. On the other hand, it is a dynamic Bayesian situation in the sense that no further information is available at the beginning and results from 5, 10, and 15 draws are known step by step. It therefore seems particularly interesting to investigate whether and, if so, how probability estimates are updated by the students based on new

information and how this is justified. We therefore asked the students to justify their decisions verbally, videotaped the interviews, and then transcribed them.

Since the students had little experience with probability scales from their mathematics lessons, their structure and use for visualizing probabilities were discussed again before the interviews in a lesson (3rd and 4th graders) or in individual discussions (5th graders) using example tasks on everyday phenomena and probability spinners. The tasks used were comparable for all students (groups) and had no connection to the Bayesian situation in the interviews.

Some Remarks on the Mathematical Background

Among other things, we are interested in to what extent students arrive at likelihood estimates that are appropriate to the information currently available. Therefore, it is necessary to first determine the relevant mathematical probabilities.

Task 1: Since one of the two bags, A or B, is selected at random at the beginning, both are equally likely. The result of the first 5 draws (E1) is very meaningful or revealing in view of the contents of the two bags, i.e., the likelihoods change significantly: $P(A|E1)=0.985$, $P(B|E1)=0.015$. In the next 5 draws (E2), more red than blue cubes are drawn, but overall only 4 of the 10 cubes drawn are red. The likelihood for bag B therefore increases, but bag A is still significantly more likely than bag B: $P(A|E1\&E2)=0.941$, $P(B|E1\&E2)=0.059$. With the next 5 draws (E3), the very significant event E1 (E3=E1) is repeated, so the likelihood that the cubes were drawn from bag A will be slightly higher after 15 draws than after 5 draws: $P(A|E1\&E2\&E3)=0.999$, $P(B|E1\&E2\&E3)=0.001$.

Task 2: Since the draws are identical, the same considerations apply as for task 1. However, because the ratios of blue:red in the two bags are less different than in task 1, the differences between the likelihoods are also slightly smaller: $P(A|E1)=0.111$, $P(B|E1)=0.889$, $P(A|E1\&E2)=0.2$, $P(B|E1\&E2)=0.8$, $P(A|E1\&E2\&E3)=0.13$, $P(B|E1\&E2\&E3)=0.97$.

The extremity of most likelihood values may seem surprising. In fact, psychological studies of similar Bayesian situations have shown a systematic tendency of observers to make overly “conservative” likelihood estimates (Peterson & Beach, 1967).

Data Analyses

In order to answer the first research question and related parts of the third research question, criteria were first derived from the mathematical background that describe the fit of the probability estimates fixed on the probability scales to the currently available information from a mathematical perspective: C1: Are *all* probability estimates updated in the appropriate direction based on the last 5 draws? C2: After 10 draws, does the difference between the probability estimates have a smaller absolute value but still the same sign as after 5 draws? C3: After 15 draws, does the difference between the probability estimates have a larger absolute value and still the same sign

as after 5 draws? C4: Are the probability estimates for bag A and bag B symmetrical to the midpoint of the probability scale? Since C1 describes whether the subject is willing and able to adjust her/his probability estimates in the appropriate direction in response to new information, it can be considered the main criterion for assessing the subject's proto-Bayesian reasoning. While C1-C3 are closely related to it; C4 arises from the specific situation and therefore plays a (subordinate) special role in evaluating the solutions. Because the selection of bags at the beginning of the tasks can be assessed differently by the students for various reasons, the prior probabilities should not be taken into account in the evaluation, also the absolute (numerical) values of the probability estimates should not play a role.

By combining the criteria, categories can be formed and the students' solutions assigned to them. If we use C*x* to symbolize fulfillment and nC*x* to symbolize non-fulfillment of the corresponding criterion, the categories we use can be described as follows and color-coded for better presentation of the results:

C1&C2&C3	C1&C2&nC3	C1&nC2&C3	C1&nC2&nC3	nC1

The inclusion of C4 doubles the number of categories; failure to fulfill this characteristic should be color-coded with a pink frame.

For example, when Maurice worked on the two tasks (cf. Fig. 1), the difference in probability estimates changed sign after 10 draws (nC2), and the probability estimates after 15 draws matched those after 5 draws (nC3). Since all estimates are symmetrical (C4), his results would be represented by yellow cells without a pink border (cf. Fig. 2, no. 3).

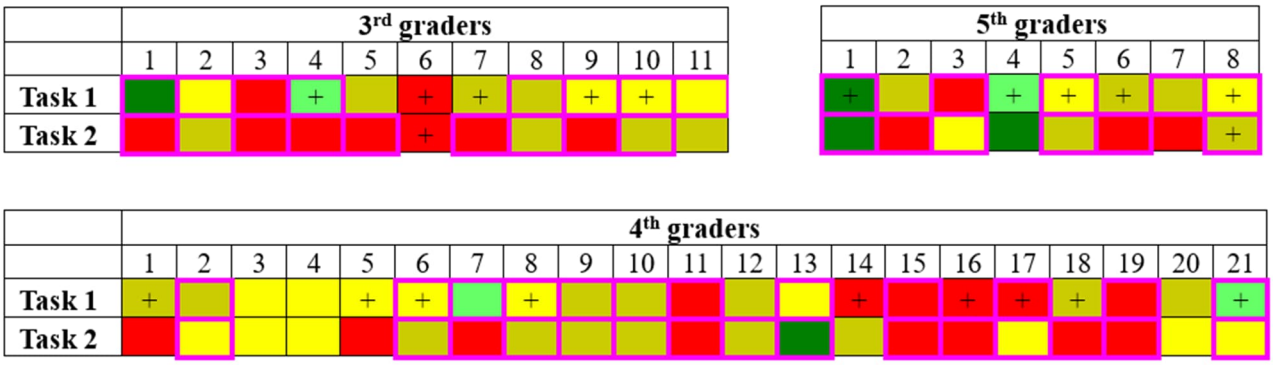
All 80 solutions were categorized independently by both authors; in cases of differing results, a second consensual categorization was performed.

SELECTED RESULTS AND DISCUSSION

Fig. 2 summarizes the results of all participating students. An additional plus sign indicates that probabilities were rated as "certain" or "impossible".

It is noticeable that in all three grades, more than half of the solutions meet criterion C1, and nC1 solutions decrease from grade to grade. In grades 3 and 4, only one solution meets the three criteria C1, C2, C3; in grade 5, there are three such solutions.

Fig. 2: Color-coded results of all 40 students



(among fewer students). This means that the students are increasingly able to respond to the last five draws with an adjustment of their probability estimates in a fundamentally appropriate manner. For 19 out of the 40 students, both solutions even meet criterion C1. However, it is also apparent that only a few cells are colored green overall; only one student in the 5th grade meets criteria C1, C2, and C3 with both solutions. This once again demonstrates the high level of difficulty of the tasks. Only seven of the 40 students meet criterion C1 with both solutions, and both were categorized in the same way. This suggests that most students do not yet have a stable approach to this type of task.

More than 65% of the cells, including 4 of 5 green colored ones, have a pink border. In these solutions, it did not matter or could not be taken into account that the probability estimates for bags A and B should be complementary to each other from a mathematical point of view.

Task 2 appears to be somewhat more difficult than task 1, but task 1 has significantly more extreme probability values. This may be due to the different ratios of blue:red in the two tasks.

Discussion

The written results indicate that students in the examined age group are already capable of appropriately adjusting their probability estimates based on new information. However, a detailed qualitative analysis of their reasoning is necessary in order to reconstruct the underlying thought processes. In particular, it seems important to examine whether the students' explanations confirm what the categorization of the written solutions suggests: on the one hand, students often succeed in updating their likelihood estimates to match the previous five draws, but on the other hand, draws further back in time are much less frequently included in their considerations.

Furthermore, it should be noted that criteria C1-C3 and the categorization based on them describe the fit of the probability estimates (and changes of them) to the currently available information from a mathematical perspective. However, the extent to which differing estimates may be consistent with each other in the Bayesian sense is less well

captured. In this regard, too, an analysis of the students' reasoning could have a wide range of potential benefits.

In summary, it can be said that (regardless of the students' justifications) the analysis of the written results indicates that even many students at the end of primary school or at the beginning of secondary school already seem to have an intuitive understanding of Bayesian situations of the second type. Probability calculation skills do not appear to be a necessary prerequisite for Bayesian reasoning. In connection with this, the tasks in the study also show that the probability scale is a suitable tool for working on more challenging probability problems and for making probability estimates representable and communicable. Overall, the results suggest that it would be entirely possible to give greater importance to Bayesian reasoning, which is so important, in secondary school mathematics classes.

References

Fornacon-Wood, I., Mistry, H., Johnson-Hart, C., Faivre-Finn, C., O'Connor, J. P. B., & Price, G. J. (2022). Understanding the differences between Bayesian and frequentist statistics. *International Journal of Radiation Oncology, Biology, Physics*, *112*(5), 1076–1082.

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.

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). The perception of probability. *Psychological Review*, *121*(1), 96–123.

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

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

Hunting, R. P. (1997). Clinical interview methods in mathematics education research and practice. *The Journal of Mathematical Behavior*, *16*(2), 145–165.

KMK. (2022). *Bildungsstandards für das Fach Mathematik Primarbereich.*

Laplace, P. S. (1902). *A philosophical essay on probabilities*. John Wiley and Sons (Original work published 1814).

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29–46.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232.