Mathematical practices when solving problems involving joint and conditional probabilities without prior instruction

José M. Rubio-Chueca[1], José M. Muñoz-Escolano[1] and Pablo Beltrán-Pellicer[1]

[1]University of Zaragoza, Faculty of Education, Zaragoza, Spain; chemarubiochueca@gmail.com

Given the significance of probability, its various interpretations, and the challenges associated with teaching it, this article presents the findings of a study that explores the strategies employed by participants when faced with a problem involving joint and conditional probabilities during the final round of the XXX Aragon Mathematical Olympiad. The onto-semiotic approach (OSA) serves as the theoretical framework for this study, with particular emphasis on the levels of algebraization. The methodology is qualitative, centred on analysing the problem statement and the participants' approaches to solving it. This allows for the observation of emerging probability meanings, the language utilized, the algebraic activity conducted, and the problem's difficulty throughout the resolution process. Our findings reveal that different levels of algebraization were observed in the resolution of each problem subsection, not necessarily correlated with the a priori difficulty.

Keywords: Probability, onto-semiotic approach, levels of algebraization.

Introduction

Nowadays, anyone should have a minimum knowledge of probability that allows them to make decisions, form judgments about relationships between events or make inferences and predictions. In this line of thought, and inspired by the notion of mathematical literacy, several authors such as Batanero, et al. (2016) stress the need for the whole population to reach a high degree of probabilistic literacy. This necessity has led to the teaching of probability in the curricula of secondary and primary schools of practically all countries. Following the model of probabilistic literacy proposed by Gal (2005), it is a matter of offering tools to the students that favour the learning of probability. From this perspective, and according to Batanero et al. (2016), special attention must be paid to problems related to the incorporation and treatment of probability and incorporate them into didactic sequences.

When designing didactic sequences, characterizing the responses of students when solving problems for which they have not received prior instruction is a very interesting research point. Teaching approaches such as problem-solving take as a starting point the prior knowledge and intuitive reasoning that students put into play when solving problems (Bingolbali & Bingolbali, 2019; Liljedahl et al., 2016). To do this, they pose problems to students before they have been explicitly explained how to solve them. Then, the teacher deploys a scaffolding using quality feedback, usually in the form of good questions and shared experiences, which provides students with opportunities to construct knowledge about the new mathematical objects that emerge when solving these problems. The characterization of the responses allows, on the one hand, to design tasks that facilitate the emergence of intuitive modes of thinking that lead to a more meaningful learning. On the other hand, it facilitates the anticipation of how the framework and management of these activities in the classroom should be. In this way, the most common mistakes can be anticipated to have small situations that enable discussing their origin (Smith & Stein, 2011).

Toh (2013) states that mathematical olympiads, beyond being punctual mathematical contests, can serve as a bank of useful resources for teachers to develop "rich" tasks with instructional purposes. At the same time, since the olympiads are activities participated by students with high mathematical ability (Olszewski-Kubilius & Lee, 2004), it is likely that a great variety of procedures and intuitive reasoning will appear in task resolutions (Leikin, 2009). In Spain, it is not usual for this type of contest to include tasks that mobilize knowledge in probability (Rubio-Chueca et al., 2021).

The purpose of this study is to characterize the mathematical practices in the responses of second grade Secondary Education (13–14-year-old) students participating in a mathematics contest. They must solve a problem involving joint and conditional probabilities without having received prior instruction. To do so, it analyses the languages used, procedures, algebraization levels, as well as errors and main difficulties in student responses.

Theoretical framework

Onto-semiotic approach to mathematical knowledge and instruction (OSA)

The OSA (Godino et al., 2019) provides tools for the analysis of teaching, of the resources involved in it and of the learning carried out by the students. This approach focuses its interest on mathematical practices giving great relevance to the notion of problem-situation (task, problem, etc.) and the mathematical objects taking place in such practices. From this perspective, Godino et al. (2019) propose an ontology of mathematical objects, as "any material or immaterial entity that intervenes in mathematical practice, supporting and regulating its realization" (p. 40). Problem-situations are tasks that induce mathematical activity while language is constituted by terms, expressions, or graphics. Procedures are operations, algorithms or techniques performed to solve a task. Concepts are given through definitions or descriptions and properties in the form of statements or propositions. Finally, arguments are used to explain or validate propositions and procedures.

The OSA also understands Elementary Algebraic Reasoning as the system of practices in the resolution of tasks in which algebraic objects and processes intervene (Godino et al., 2014). Four levels of elementary algebraization are proposed depending on the degree of generality and formalism involved that characterize them. The application of the levels proposed by Burgos, Batanero, et al. (2022) provides us with criteria to distinguish categories of meanings in the progressive construction of our student's probability reasoning. The arithmetic level or level 0 is characterized by the absence of algebraic reasoning and by the appearance of the first-degree generality objects. For example, arithmetic operations on natural numbers associated with procedures such as distinguishing and counting possible and favourable cases and calculating the ratio between favourable and possible cases. The proto-algebraic level is formed by level 1 (incipient) and level 2 (intermediate). In level 1 the first properties of operations with rational numbers are recognized, the concept of equivalence is recognized too, and the first symbols appear without operating with them. In terms of functionality at this level 1, Laplace's Rule (simple probability), Sum Rule (compound probability) and Product Rule (compound probability) are identified. At level 2, symbolic representations are used to express intensive objects related to the problem's context. Equations of the form $Ax+B=C$ are solved that may appear. In functional tasks a general rule is recognized. The use of different diagrammatic or tabular languages (tree diagram, probability table, contingency table) can act as an icon of a structure of relationships that allows obtaining knowledge related to the probability distribution.

Errors and degrees of difficulty in conditional probability problems

Problems regarding conditional probability are not an easy task for students. Various works emphasize the difficulty of its realization (Díaz & Batanero, 2009; Díaz et al., 2010). In this regard, Huerta et al. (2016) address the study of the difficulties and errors made by students when solving this type of problems. They provide a series of variables that allow them to code the different responses given by students where three distinct types are distinguished: variables from approach, process and result. Regarding result variables, they "allow us to observe and analyse what the solver declares as an answer to the question posed by the problem" (Huerta et al., 2016, p. 342). Among these result, four variables are distinguished: result (if the solver provides any answer to the problem of any kind regardless of its suitability), number (if the solver gives a correct numerical answer or has some arithmetic error), correct descriptor (if numerical interpretation is correct using a verbal or symbolic description own from probability) and incorrect one (if a verbal or symbolic description of probability is provided that does not provide desired meaning).

Huerta and Arnau (2017) also extend these variables to the study of joint and total probability problems. Through these variables, they quantify a posteriori the difficulty of some problems and establish the following indices: of perceived difficulty (measure of whether the problem is perceived as "difficult" to be solved), difficulty of the problem (measure of the difficulty to give an answer once it is addressed resolution), difficulty of solving the problem (measure of the difficulty if the numerical solution is correct once an answer is obtained), difficulty in describing the response to the problem (measure indicating the number of solvers who give a correct description with respect to those who give an answer) and difficulty in giving a correct description (measure indicating the difficulty in giving a correct description taking into account all descriptions).

Methods

The main objective of this research is to analyse the resolutions given by students who have not received prior instruction on a total and conditional probability problem. Content analysis is used for the analysis of mathematical objects related to probability, performing an exploratory-descriptive analysis. This analysis is supported by the categories and tools of OSA and by the coding of variables that consider difficulty. To carry out the work, we identify and select the mathematical practices present in the written resolutions to a problem on probability. We code specific examples according to the categories that allow us to carry out a study of qualitative data using MAXQDA software. For greater coherence in the process, this is reviewed through triangulation analysis by researchers. To this end, several meetings were arranged by the three researchers to analyse the students' responses. After several discussions in the review of the data, it was decided to analyse the mathematical objects and difficulties present in the study, due to the interest in being part of a problem-situation of a sequence aimed at conditional probability.

The mathematical objects involved in the solvers' responses are analysed with some of the categories provided by the OSA (Godino et al., 2019), specifically the procedures and languages used in solving the two questions that appear in the problem. Some levels of algebraization are also analysed (Burgos, Batanero, et al., 2022). The variables of the results of Huerta et al. (2016) and Huerta and Arnau (2017) will be considered, defining the following levels of difficulty measured in percentages: Appreciated Difficulty, as the ratio between the number of solvers who address the problem and the total number of solvers to whom its resolution was proposed; Problem Difficulty, as the ratio between those who do not respond to the problem and those who address its solution; Difficulty of Problem Solution, through the ratio between those who give an incorrect numerical response to the problem and those who address it; Difficulty for Description of Problem Response, as the ratio between those giving a description and those giving a response to it; Difficulty for Correct Description of Problem Solution, as the ratio between correct descriptions given and total descriptions given.

The sample consists of 108 secondary school students aged between 13 and 14 who participated in the final phase of the XXX Mathematics Olympiad in Aragon (Spain) in 2022. The units of analysis are the 108 solutions by these participants to the following problem:

> Twins Alba Pasol and Carmen Pasol are basketball players who are indistinguishable at first sight for their coach. Alba's field goal accuracy is 40%, while Carmen's is 10%. Alba's jersey number is 9 and Carmen's is 2. During halftime, they switched jerseys completely at random. With only a few seconds left in the game, their team is losing by one point and there's only one shot at the basket. The coach chooses to bring out the player wearing the 9 jersey to shoot, leaving the other on the bench. a) What is the probability of winning the game? Why?b) The player scores and they win. What is the probability that it was Alba who scored? Why?

Question a) corresponds to a problem on total probability. It is asking for a simple probability which can be obtained by summing two joint probabilities considering the independence of events and the product rule. Question b) can be associated with a problem on conditional probability involving the Bayes theorem, since it requests the calculation of posterior conditional probability. However, it is not necessary to use the theorem in a formal way since it is possible to answer both questions with other procedures. For example, one can arrive at an answer for b) taking into account the restriction that conditional probability imposes on sample space and using Laplace's probability.

Results

Of the 108 students that tackled the problem, 45 gave a correct answer to the probability of winning the game by following suitable strategies (Table 1). To reach that solution, 25 did so using a classical meaning of probability by calculating the total probability as the sum of joint probabilities. Other students, also from the classical meaning of probability, opted for more intuitive and less formal procedures. Specifically, 17 students gave a correct answer using as procedure an arithmetic mean of conditional probabilities assuming implicitly or explicitly equiprobability of two elementary and exclusive events A = {Carmen wears shirt with number 9} and B = {Alba wears shirt with number 9}, and 3 students did so by considering the concept of ratio between percentages assuming the sum of probabilities of successes as partial cases or partial realizations.

Table 1: Number of participants who used each of the correct strategies in part a)

Strategies used in section a)	Number of participants
The use of the concept of ratio between percentages (partial)	3
The use of the rule of sum of probabilities	25
The use of arithmetic mean in conditional probabilities	17

As shown in Table 2, of the 106 students who attempted the second part of the problem to answer the posterior probability, 30 provided correct answers using a classical meaning, with 23 establishing some proportion regarding conditional or joint probabilities, and 7 used the mathematical concept of ratio between percentages assuming the sum of probabilities of successes as total cases or total realizations. As expected, none used Bayes' theorem for question b), as it falls outside the official curriculum for the participants' age group. Beyond procedures seen in Tables 1 and 2, solvers employed auxiliary procedures such as using equiprobability, determining sample space, or calculating joint probability through the product of probabilities in independent events.

Table 2: Number of participants who used each of the correct strategies in part b)

Strategies used in section b)	Number of participants
The use of the concept of ratio between percentages (total)	7
The use of proportion in Probabilities	23

To determine the algebraic level of reasoning carried out by the students in the correct answers, we consider the diverse nature of mathematical objects and processes involved in operational and discursive practices. In answer a) the highest percentage, 55%, is associated with Level 1, corresponding to calculation of total probability, considering equiprobability in the choice of shirt and joint probability. In b), all correct answers are associated with Level 1, corresponding to most as the use of proportion in conditioned probabilities and knowledge of percentages (Table 3).

Table 3: Levels of algebraization in the correct responses

Levels of Algebraization	Answer a)	Answer b)
Arithmetic level or Level 0	15	0
Proto-algebraic level	30	30
Level 1	26	30
Level 2	4	0

Regarding the difficulty levels observed in the problem (Table 4), all students attempted both questions posed, except for two participants who did not do so in part b). It is worth noting that all those who attempted the problem did not abandon it midway and ultimately reached some form of solution. Of these, 17.9% gave a correct numerical and descriptive solution to the problem. On the other hand, 50% found difficulty in finding the solution to question a), with 32.4% making errors when tackling it by using total probability. The main errors made by students were confusing the total probability with: the mean of joint probabilities (8.3%), joint probability (6.5%), and the sum of conditional probabilities (3.7%) mostly corresponding to Level 1. Conversely, 67% of participants encountered difficulties in question b), with 59.4% making errors when establishing posterior probability using incorrect procedures or strategies unrelated to proportions, incorrectly associating probabilities and highlighting main confusions where generalizations were used (Level 1) and joint probability was equated with conditional probability (18.9%) or initial probability with conditional probability (17.9%). In these cases, one should address this task by using Bayes' theorem which corresponds with Level 2.

Table 4: Identified difficulties in the problem

Types of difficulty	Answer a)	Answer b)
Appreciated difficulty	100%	98%
Problem difficulty	0%	0%
Difficulty of problem solution	50%	67%
Difficulty for description of problem response	87%	92%
Difficulty for correct description of problem solution	41%	41%

Regarding the registers and linguistic representations used by participants, we first note imprecisions in verbal register usage, with 26.7% confusing "possibility" with "probability". In the symbolic-conjunctist register, 2.7% used functional notation and 11.1% employed letter symbols. As the problem statement uses percentages, it is expected that most answers (95.4%) would feature them. However, fractions (45.4%), decimals (20.4%), and natural numbers (22.2%) also appeared. In the diagrammatic register, tree diagrams were observed in 27.8% of responses, but their use was not significant for solving problem parts. Only 15.6% and 16.7% of participants who correctly solved parts a) and b), respectively, incorporated tree diagrams in their answers.

Conclusions

Despite the problem-situation's potential unfamiliarity or unrealistic nature, the numerical language used (percentages) is common in its context (sports-basketball). The problem also involves familiar auxiliary procedures, prompting students to attempt solving it without prior knowledge of specific methods. However, less than a fifth provided accurate numerical descriptions and answers for both parts. Difficulty percentages were similar for both parts, with the second part seemingly more complex due to the potential necessity of Bayes' theorem from a classical probability perspective.

Various correct strategies were identified in students' responses, such as employing total probability for part a) and using proportion in probabilities for part b). To be accurate, the latter requires restricting the sample space to apply conditional probability. The diversity and correctness of many procedures are attributed to conditions like equiprobability in choosing two players. After addressing this task, it would be optimal to introduce a similar problem without equiprobability, within a problem-solving approach (Bingolbali & Bingolbali, 2019; Liljedahl et al., 2016).

Students used a range of registers and linguistic representations, with some exhibiting terminological confusion. Although the problem statement referred only to percentages, many students naturally translated this into other registers (natural numbers, decimals, and fractions). The use of diagrammatic registers, such as tree diagrams, was less prevalent, suggesting a lack of intuitiveness. Thus, introducing such registers in the classroom is crucial (Roldán, et al., 2018).

For the problem's first part, correct answers mainly involve the initial three algebraization levels. Level 0 corresponds to considering conditional probability means, Level 1 to using total probability, and Level 2 to identifying sample space and probabilities. For the second part, all correct answers fall under Level 1, focusing on establishing proportions using ratios, percentage knowledge, and probability properties. Some students establish proportions involving two fractions or employ Laplace's rule with rational numbers, both characteristic of Level 1.

In conclusion, the sample of Mathematics Olympiad finalists allows for uncovering diverse approaches, as such contests attract high-ability students. Future studies should consider broader student populations, comparing procedures and reasoning to develop didactic sequences based on "teaching through problem solving" (English & Sriraman, 2010). Our study informs instructional design by highlighting informal strategies and difficulties with conditional probability, promoting high didactic suitability (Godino et al., 2023). The task also connects different probability meanings (Batanero et al., 2016) with proportional reasoning in probability (Burgos, Albanese, et al., 2022).

Acknowledgements

This project (PID2019-105601GB-I00) was funded by MCIN/ AEI /10.13039/501100011033 and supported by the S60_23R group, funded by the Government of Aragón.

References

Batanero, C., Chernoff, E. J., Engel, J., Lee, H. S., & Sánchez, E. (2016). *Research on teaching and learning probability*. Springer Nature. https://doi.org/10.1007/978-3-319-31625-3_1

Bingolbali, F., & Bingolbali, E. (2019). One curriculum and two textbooks: opportunity to learn in terms of mathematical problem solving. *Mathematics Education Research Journal, 31*, 237–257.

Burgos, M., Albanese, V., & López-Martín, M. M. (2022). Prospective primary school teachers' recognition of proportional reasoning in pupils' solution to probability comparison tasks. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.) *Proceedings of the CERME12* (pp. 837–844). Free University of Bozen-Bolzano and ERME.

Burgos, M., Batanero, C., & Godino, J. D. (2022). Algebraization levels in the study of probability. *Mathematics, 10*, 91. https://doi.org/10.3390/math10010091

Díaz, C., & Batanero, C. (2009). Students' formal knowledge and biases in conditional probability reasoning. Do they improve with instruction? *International Electronic Journal of Mathematics Education, 4*(3), 131–162. https://doi.org/10.29333/iejme/234

Díaz, C., Batanero, C., & Contreras, J. M. (2010). Teaching independence and conditional probability. *Boletín de Estadística e Investigación Operativa, 26*(2), 149–162.

English, L., & Sriraman, B. (2010). Problem solving for the 21st century. In L. English & B. Sriraman (Eds.), *Theories of mathematics education* (pp. 263–290). Springer.

Gal, I. (2005). Towards "probability literacy" for all citizens: Building blocks and instructional dilemmas. In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning*, (pp. 39–63). Springer. https://doi.org/10.1007/0-387-24530-8_3

Godino, J. D., Aké, L. P., Gonzato, M., & Wilhelmi, M. R. (2014). Niveles de algebrización de la actividad matemática escolar. Implicaciones para la formación de maestros [Algebrization levels of school mathematics activity. Implication for primary school teacher education]. *Enseñanza de las Ciencias, 32*(1), 199–219. http://dx.doi.org/10.5565/rev/ensciencias.965

Godino, J. D., Batanero, C., & Font, V. (2019). The onto-semiotic approach: implications for the prescriptive character of didactics. *For the Learning of Mathematics, 39*(1), 37–42.

Godino, J. D., Batanero, C., & Burgos, M. (2023). Theory of didactical suitability: An enlarged view of the quality of mathematics instruction. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(6), em2270. https://doi.org/10.29333/ejmste/13187

Huerta, M. P., & Arnau, J. (2017). La probabilidad condicional y la probabilidad conjunta en la resolución de problemas de probabilidad [Conditional and joint probabilities in conditional probability problem solving]. *AIEM, 11*, 87–106.

Huerta, M. P., Edo, P., Amorós, R., & Arnau, J. (2016). Un esquema de codificación para el análisis de las resoluciones de los problemas de probabilidad condicional [Encoding scheme for the analysis of resolutions of conditional probability problems]. *RELIME, 19*(3), 335–362.

Leikin, R. (2009). Exploring mathematical creativity using multiple solution tasks. In R. Leikin, A. Berman, & B. Koichu (Eds.), *Creativity in mathematics and the education of gifted students* (pp. 129–145). Sense Publisher. http://dx.doi.org/10.1163/9789087909352_010

Liljedahl, P., Santos-Trigo, M., Malaspina, U., & Bruder, R. (2016). *Problem solving in mathematics education*. Springer. https://doi.org/10.1007/978-3-319-40730-2

Olszewski-Kubilius, P., & Lee, S. (2004). The role of participation in in-school and outside-of-school activities in the talent development of gifted students. *Journal of Secondary Gifted Education, 15*(3), 107–123. http://dx.doi.org/10.4219/jsge-2004-454

Roldán López de Hierro, A. F., Batanero, C., & Beltrán-Pellicer, P. (2018). El diagrama de árbol: un recurso intuitivo en Probabilidad y Combinatoria [Tree diagram: an intuitive resource in probability and combinatorics]. *Épsilon, 100*, 49–63.

Rubio-Chueca, J. M., Muñoz-Escolano, J. M., & Beltrán-Pellicer, P. (2021). La probabilidad en los problemas de olimpiadas matemáticas de Secundaria en España [The probability in the problems of high school math olympics in Spain]. *Contextos Educativos*, *28*, 29–50.

Smith, M. S., & Stein, M. K. (2011). *5 Practices for Orchestrating Productive Mathematics Discussions*. NCTM.

Toh, T. L. (2013). Mathematics Competition Questions and Mathematical Tasks for Instructional Use. In B. Kaur (Eds.), *Nurturing Reflective Learners in Mathematics: Yearbook 2013* (pp. 189–207). World Scientific, AME.