THE 'SAME' PROBLEM IN THREE PRESENTATION FORMATS: DIFFERENT PERCENTAGES OF SUCCESS AND THINKING PROCESSES

M. Pedro Huerta (1) Mª Ángeles Lonjedo (2)

(1) Universidad de València, Spain, (2) IES Montserrat, Spain

ABSTRACT

In this paper we investigate the influence that presentation format of a conditional probability problem has on students' problem solving behavior. We not only focus on the way numerical data is presented but also on information in text form that refers to a conditional probability. We report that students' behavior changes depending on data presentation, and the percentage of students that succeed in solving a problem increases if we change the presentation format of the problem in a suitable way. We conclude by making some proposals for teaching conditional probability problem solving.

INTRODUCTION

From a psychological point of view, some authors (eg, Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996; as opposed to Evans, Handley; Perham, Over & Thompson, 2000; Girotto & Gonzalez, 2001), suggest that people calculate Bayesian inferences better if the information is expressed in terms of frequencies rather than probabilities. They report that participants perform better because there is a strong relationship between data format and the rules required to answer the problem. In (natural) frequency formats, these rules are less complex than in probability formats and can facilitate reasoning in complex Bayesian situations. Other authors argue that the reason doesn't lie in data format but in information structure and the form of the question, making the problem much easier to understand.

Almost all the conditional probability problems used in the research mentioned above were structurally isomorphic. Usually these problems were considered in pairs. Apart from data format, all of them can be mathematically (or symbolically) read as follows: Known $p(E)$, $p(+|E)$ and $p(+|\sim E)$ calculates $p(E|+)$, otherwise known as the Disease Problem. All of these problems we label by means of a vector (1,0,2) (see Carles & Huerta, 2007) that indicates the number and type of known numerical data we have in text of problem: in this case, we use one absolute probability and two conditional probabilities to calculate the unknown numerical data. Number 0 means that we do not know the intersection probabilities. In order to calculate $p(E|+)$ we need Bayes' rule and the Theorem of Total Probability. But these problems are different (ie, not isomorphic) if we consider that problem information is presented using different data formats: frequencies in one case and percentages in the other. None of them used probabilities (numbers in [0,1]) as a data format.

On the other hand, the subjects performing problems in these studies weren't considered to be math students, with some knowledge of making Bayesian inferences, but as people naïve in this topic and in probabilistic reasoning. For this reason, the main objective of these investigations was to explore reasons for successful problem solving according to data format and discovering which data format facilitates the most success.

From the point of view of the didactic of mathematics, in a previous paper (Huerta & Lonjedo, 2006) we showed that different presentation formats of a conditional probability problem resulted in different student problem solving behavior. This paper investigated the processes involved in solving 16 conditional probability problems. These paired problems were structurally isomorphic but used different data formats. One of the conclusions relates to differences in student behavior in solving the 'same' problem when data is expressed in terms of percentages as opposed to probabilities; ie, if data is expressed in terms of percentages, then students usually solve these problems using mainly arithmetical thinking strategies, whereas if data is expressed in terms of probabilities, one can recognize probabilistic thinking strategies in solving these problems.

In CERME4 (Huerta & Lonjedo, 2006) we presented a report highlighting the problem solving processes when data is expressed in terms of percentages and probabilities. In CERME5, however, we present a piece of work that focuses on the process of solving three problems that are isomorphic in structure but where the information is expressed in three different formats: in terms of percentages, probabilities and absolute frequencies. Thus, the aims of this paper are: (1) to study the influence of data format in conditional probability problems on students' behaviors and success; (2) to study the influence of semantic and syntactic aspects on the students' success in solving these problems.

THE RESEARCH PROBLEM

Let us consider problems P7 and P15 (see Table 1) used in Huerta & Lonjedo (2006). These problems can be mathematically read as follows: Known $p(A)$, $p(B)$ and $p(A|B)$, calculate $p(B|A)$. It is a (2,0,1) problem asking for an inverse probability from a known conditional probability. The Bayes' rule solves this problem.

We know that in general, the percentage of students that were successful in solving both problems was very low (Huerta & Lonjedo, 2006). Only math college's students were successful whereas secondary school students were not. Therefore, we considered a new problem, P1, structurally isomorphic with both P7 and P15, but with the data expressed in a different format. From the experience we had with problems P7 and P15, we also reconsidered the information in text form that was used to describe the data as conditional probabilities and expressed them with the same grammatical structure, in order to avoid as many misunderstandings as possible. One of these misunderstandings, for example, relates to students' confusion between the conditional probability and the intersection probability. Quantities in P1 are always absolute frequencies, except when referring to a conditional probability, in which case percentages must be used. Moreover, in P1 we also tried to avoid both semantic difficulties, using for example a more understandable sentence for the conditional probability, and semiotic difficulties, writing data for the absolute probabilities in terms of absolute (natural) frequencies. In Table 1 we can see the three problems we are referring to which are, essentially, three versions of the 'same' problem.

P7 PROBLEM Data in percentages	*60% of students in a school succeeded in Philosophy and 70% in Mathematics. Moreover, 80% of the students that succeeded in Mathematics also succeeded in Philosophy. If Juan succeeded in Philosophy, what is the probability that he also succeeded in Mathematics?*
P15 PROBLEM Data in probabilities	*In a school, the probability of success in Philosophy is 0.6 and in Mathematics, 0.7. Choosing a student at random among those that succeeded in Mathematics, the probability that he/she also succeeded in Philosophy is 0.8. If Juan succeeded in Philosophy, what is the probability that he also succeeded in Mathematics?*
P1 PROBLEM Data in frequencies	*In a class of 100 students, 60 succeeded in Philosophy and 70 succeeded in Mathematics. Among those who succeeded in Mathematics, 80% also succeeded in philosophy. Of those who succeeded in Philosophy, what percentage of students also succeeded in Mathematics?*

Table 1. Three versions of the 'same' problem

These problems form part of a broader research that tries to investigate the processes involved in solving conditional probability problems. One of the questions we try to answer has to do with the relationship between data presentation format and students' problem solving process.

METHOD

All three problems were items in a test administered to students of different ages and mathematical ability: Lower secondary school (13-14 year olds), upper secondary school (15-18 year olds) and 2nd year math students at university. In table 2 we can see the distribution of the student sample involved in this research. Only students from upper secondary school and university were taught about conditional probability, whereas students from lower secondary school were not.

School Level	P7	P15	P1
Lower Secondary School	11	5	31
Upper Secondary School	52	26	39
University (Math College)	4	2	10
Total	67	33	80

Table 2: Number of students that tried to solve each problem

Because we were not only interested in students' success in solving problems but also in resolution processes, successful or not, we designed a set of descriptors to analyze problem solving behaviour, as follows:

1. - Problem solving process with success. This descriptor reports students' successful behaviors in finding the correct result to the problem. Depending on the different reasoning shown by students during the problem solving process, we distinguish:

1.1. Problem solving processes that include a type of thinking that is exclusively arithmetical: Students think in quantities and not in events and their probabilities, at least in a conscious way.

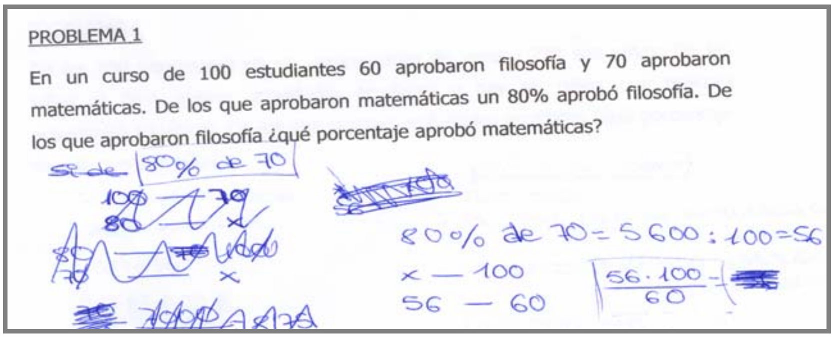
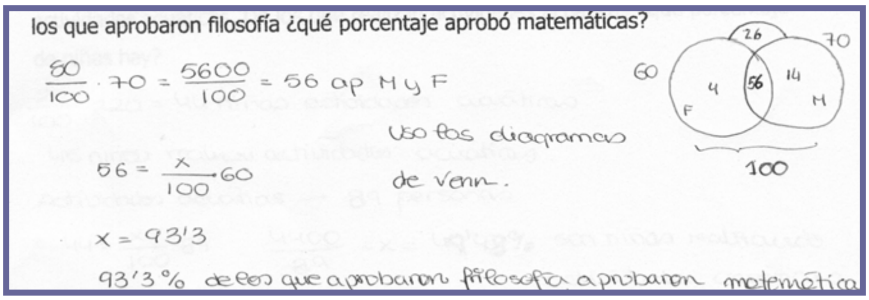


Figure 1: Example of thinking process (in P1) classified in 1.1, exclusively arithmetical thinking.

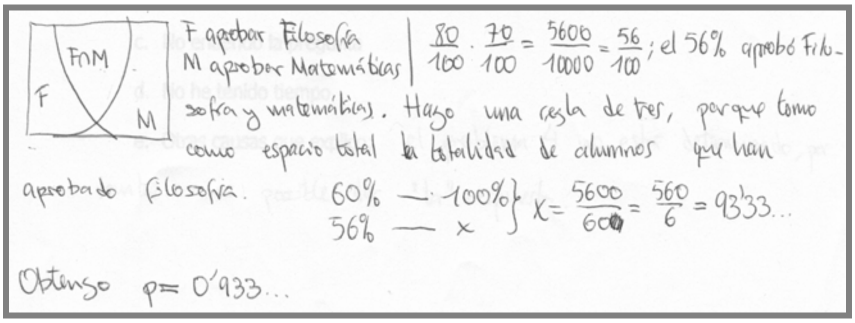
1.2. Problem solving processes that include a type of thinking that is mostly arithmetical: Students think in quantities but they recognize events and their associated frequencies or percentages.



F to succeed in Philosophy, M to succeed in Mathematics. 56 succeeded in M and F, I use the Venn diagrams, 93.3% among those successful in Philosophy succeeded in Mathematics

Figure 2: Example of thinking process (in P1) classified in 1.2, mostly arithmetical thinking.

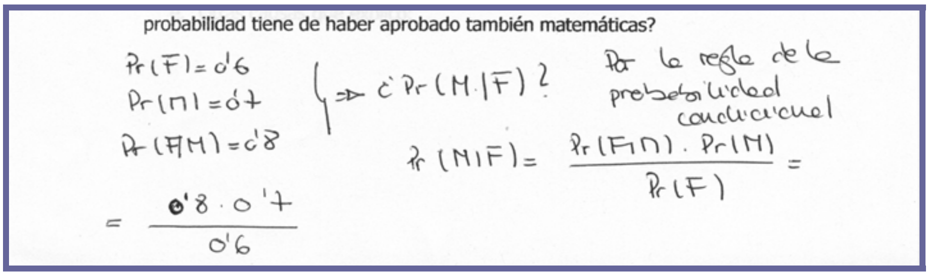
1.3. Problem solving processes that include a type of thinking that is basically probabilistic. In solving problems, students think arithmetically in quantities. These quantities are not used explicitly as probabilities. However, students recognize events and assign probabilities to them without using probability rules in a conscious way.



F to succeed in Philosophy, M to succeed in Mathematics … 56% succeeded in Philosophy and Mathematics. I do a rule of three because I get as total space the totality of students that succeeded in Philosophy ….I get p= 0.933..

Figure 3: Example of thinking process (in P7) classified in 1.3, basically probabilistic thinking.

1.4 Problem solving processes that include a type of thinking that is exclusively probabilistic. Students recognize events, assign probabilities to the events and explicitly use probability rules in order to find the result of the problem.



Because the rule of conditional probability

Figure 4: Example of thinking process (in P15) classified in 1.4, exclusively probabilistic thinking.

2. - Problem solving processes without success. This descriptor reports students' behaviors that were unsuccessful. Within this general descriptor, we consider a more specific descriptor that describes students' semantic and syntactic difficulties, misunderstandings and mistakes, as follows:

2.1 Difficulties. We analyze the solvers' difficulties related to semantic and semiotic variables.

2.1.1 Semantic Difficulties. We analyze grammatical structures in descriptions used to express conditionality both as data (known and unknown) and as text.

2.1.2 Syntactic Difficulties. We analyze formats of data and question presentation in problems.

2.2 Mistakes. We analyze students' mistakes related to difficulties.

2.2.1. Mistakes as a result of semantic difficulties. These mistakes are undesirable interpretations of data in problems when students are translating them from usual language into symbolic language. Sometimes problem solving processes are coherent with students' interpretations. These mistakes could appear early in the process, both in recognizing events and in assigning probabilities to events.

2.2.1.1. Students' interpretations of conditional probability when it is data. We can distinguish the following interpretations:

2.2.1.1.1. Interpretation of the conditionality as an intersection event (Ojeda, 1995).

2.2.1.1.2. Interpretation of the conditionality as an absolute probability. Student answers the question $p(A|B)$ by means of $p(A)$.

2.2.1.2. Students' interpretations of the conditional probability when it is a question.

2.2.1.2.1. Interpretation of the conditionality as an intersection event. Students answer questions about a conditional probability by means of an intersection probability.

2.2.1.2.2. Interpretation of the conditionality in question as conditionality in data. Student interprets $p(A|B)$ as equal to $p(B|A)$, the first being probability in data and the second, the question.

3. - Others. We place here all students' resolutions that are impossible to be qualified or explained by the other descriptors. These include blank answers, answers without workings, unrecognizable signs etc.

There is another source of mistakes in solving these problems that we did not consider in this work. This concerns the misuse of decimal numbers, percentages, formulas and mathematical calculations. These mistakes, of course, would hinder successful problem resolution.

SOME RESULTS

The percentage of students that succeeded with problems P7 and P15 was 6%. All of them were students from University. However, the percentage of students that succeeded with P1 was 36.25%. This figure includes students from all age levels.

From the data in Table 3, we can see that the percentage of students that did not try to solve the problems decreases in order, from the highest (66.67%) when the data is in the form of probabilities, to the lowest (21.25%) when it is presented as frequencies. Consequently, there is an appreciable increase in the percentage of students that succeeded in solving the problems, from the lowest when the data is in the form of probabilities and percentages (5.97% and 6.06% respectively), to the highest (36.25%) when presented as frequencies. In other words, the success rate when data was presented as percentages or probabilties was extremely low compared with the number of students that succeeded with the data in the form of frequencies. We can explain these differences using the descriptors that classify students' mistakes in the problem solving process. There is a very high percentage (89.74%) of students that did not succeed in P7 because they incorrectly interpreted conditionality data either as a question or as known data. Similar mistakes occurred in P15. However, this misunderstanding occured much less (20.6%) with data in the form of frequencies. A very high percentage (89.65%) of students that were successful when the problem was described using frequencies used arithmetical reasoning. No one used this type of reasoning when problem solving with percentages or probabilities. Only 3 students out of the 29 that succeeded with the problem in frequencies used probabilistic reasoning.

DISCUSSION

The presentation of data in a conditional probability problem has some influence on the students' success and problem solving behavior (see Table 3). We think that the increase in students' success (Table 3) is due to two factors: avoiding words that provoke ambiguity; and presenting data as absolute frequencies. For example, avoiding words like *y* (and) or *también* (also) in problem descriptions prevented students from confusing conditional probability with intersection probability. The expression *De los que* (Among those who), that refers to conditionality both as data and as question, improved students' interpretation.

When data is expressed in terms of absolute frequencies and conditional probability as a percentage (Lonjedo Huerta (2006), p. 531), we believe that the chances of successful problem resolution are enhanced with a consequent increase in the percentage of successful students. Gigerenzer (1994) reports that in order to solve probability problems, our minds are better equipped if all data is expressed in terms of frequencies. We agree, although when referring to a conditional probability problem, we would like to add that if one of the data is a conditional probability, then it must be expressed in terms of a percentage in order to differentiate it from other data expressed in terms of absolute frequencies. In this manner, we can help students to interpret conditionality correctly.

Descriptors	P7 Problem Data in percentages	P15 Problem Data in probabilities	P1 Problem Data in frequencies
1. Problem solving process with success	(4 out of 67) 5.97%	(2 out of 33) 6.06%	(29 out of 80) 36.25%
1.1 Thinking is exclusively arithmetical	0	0	(7/29) 24.13%
1.2 Thinking is mostly arithmetical	0	0	(19/29) 65.52%
1.3 Thinking is basically probabilistic	(1/4) 25%	0	0
1.4 Thinking is exclusively probabilistic	(3/4) 75%	(2/2) 100%	(3/29) 10.34%
2. Problem solving processes without success	(39 out of 67) 58.21%	(9 out of 33) 27.27%	(34 out of 80) 42.50%
2.2.1.1.1. Interpretation of the conditionality as an intersection event	(13/39) 33.33%	(4/9) 44.44%	(1/34) 2.94%,
2.2.1.1.2. Interpretation of the conditionality as an absolute probability	0	(1/9) 11.11%,	0
2.2.1.2.1. Interpretation of the conditionality in question as an intersection event	(22/39) 56.41%	(4/9) 44.44%	(6/34) 17.65%
2.2.1.2.2. Interpretation of the conditionality in question as conditionality in data	(3/39) 7.69%	0	(2/34) 5.8%,
3. Others	(24 out of 67) 35.82%	(22 out of 33) 66.67%	(17 out of 80) 21.25%

Table 3. Percentages of students in relation to the descriptors.

Of those students who succeeded, arithmetical thinking strategies were typically used in problem solving, although they also demonstrated recognition of events and their frequencies and percentages. However, the thinking process used in solving the problems seems to be strongly related to data format. Students that succeeded with the problem described in frequencies used mainly arithmetical reasoning, whereas those who were successful with the problem expressed as percentages or probabilities used probabilistic reasoning. In general, when students were solving these problems they use the data explicitly mentioned without translation from one format to another. Only in a few cases (3 out of 29) did students translate frequencies into probabilities in order to solve the problem (P1) using probabilistic reasoning. These were all University students.

CONCLUSION

The concept of natural frequencies introduced by Gigerenzer and his colleagues (a good discussion about this concept can be read in Hoffrage, Gigerenzer, Krauss & Martignon (2002)), has produced some proposals about natural frequencies-based teaching in the last two ICOTS. Martignon & Wassner (2002) proposed using (natural) frequency trees in order to read quantities in a typical Bayesian problem and facilitate secondary school students problem solving. They concluded that students trained with frequency trees performed significantly better than students formula-trained. In a similar way, but in a computer environment, Sedlmeir (2002) proposed using frequency trees and frequency grids to help students read (1,0,2) problems with data in percentages using an isomorphic problem with data in (natural) frequencies. Moreover, Martignon & Kurz-Milcke (2006) proposed a method of training younger students by means of arithmetic urns and tinker-cubes in stochastic reasoning.

This work also has implications for teaching problem solving in conditional probability. In agreement with the authors mentioned above, we propose organizing the process of teaching this topic by starting with solving problems like P1, prior to attempting problems like P7, and then finally moving on to problems like P15. Our reasoning consists of introducing students to the subject by means of rates and proportions, making Bayesian inferences with data in (natural) frequencies and using arithmetic reasoning, followed by percentages employing basically probabilistic reasoning, and finally by means of probabilities with an exclusively probabilistic reasoning approach. (0,0,3) problems, like the P4 problem that was presented by Carles & Huerta (2007) in CERME5, are not able to be solved using only arithmetic reasoning and, consequently, by means of (natural) frequencies. They require probabilities and probabilistic reasoning strategies.

Finally, we also agree with other authors (eg, Ojeda, 1995; Girotto & Gonzalez, 2001) that one of the main sources of student error involves misinterpretation of the conditionality and the probability of the intersection event. The incidence of this misinterpretation could be reduced if we teach conditional probability problem solving in the way we propose in this work.

REFERENCES

Carles, M.; Huerta, M. P. (2007). *Conditional probability problems and contexts: The Test of Diagnostic Context*. Communication presented to the

Stochastic Working Group at CERME V. Larnaca, Cyprus. February 2007.

Cosmides, L.; Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* 58, pp. 1-73.

Evans J.St.B.T.; Simon, J.H.; Perham, D.E.O; Thompson, V.A. (2000). Frequency versus probability formats in statistical word problem. *Cognition* 77, pp. 197-213.

Gigerenzer, G. (1994): Why the distinction between single-event probabilities and frequencies is important for psychology (and vice-versa), In G. Wright y P. Ayton (Eds.) *Subjective probability*, pp. 129-161, Wiley

Gigerenzer, G.; Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency format. *Psychological Review*, 102, pp. 684-704.

Girotto, V.; Gonzalez, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition* 78, pp. 247-276.

Hoffrage, U.; Gigerenzer, G.; Krauss, S.; Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, 84, pp. 343-352.

Huerta, M.P; Lonjedo, M. A. (2006) The Nature of the quantities in a conditional probability problem. Its influence in the problem solving behavior. In M. Bosch (ed) (2006), *European Research in Mathematics Education IV. Proceedings of the Four Congress of the European Society for Research in Mathematics Education*, pp. 528-538. Sant Feliu de Guíxols, Spain.

Martignon, L., Wassner, Ch. (2002). *Teaching Decision Making and Statistical Thinking with Natural Frequencies*, in http://www.stat.auckland.ac.nz/publications/1/10_52_ma.pdf

Martignon, L.; Kurz-Milcke, E. (2006). *Educating Children in Stochastic Modeling: Games with Stochastic Urns and Colored Tinker-Cubes*, in http://www.stat.auckland.ac.nz/publications/17/C443.pdf

Ojeda Salazar, A.M. (1995). Dificultades del alumnado respecto a la probabilidad condicional, *UNO Revista de Didáctica de las Matemáticas*, nº5, julio 1995, pp. 37-44

Sedlmeier, P. (2002). *Improving Statistical Reasoning by Using the Right Representational Format*, in http://www.stat.auckland.ac.nz/publications/1/3f3_sedl.pdf