# Assessing Difficulties of Conditional Probability Problems[1]

M. Pedro Huerta, Fernando Cerdán, Mª Ángeles Lonjedo, Patricia Edo

Universitat de València (Spain)

## Abstract

*In this paper we define some difficulties of conditional probability problems and their measurements. We also show some results about the measurements of those difficulties in relation to the structure of problems, the contexts in which they are formulated and the sample of students that solved them. Finally, we also comment on the possible influences of these factors on the difficulties of problems.*

## Introduction

Some research reports inform us of the difficulties and the low success rate students have in solving tasks or problems involving Bayes' rule, or in general in solving conditional probability problems (Diaz & De la Fuente, 2007). These difficulties, it is said, depend on multiple factors. One of them is the cognitive complexity of the concept. Sometime this complexity is related to data format and the presentation of the data in problems (Estrada & Diaz, 2006; Huerta & Lonjedo, 2006; Lonjedo & Huerta, 2007), the context in which problems are put forward (Ojeda, 1996: Carles & Huerta, 2007) and the particular language with which conditionality and events are expressed (Maury, 1984). However, in general, these studies do not pay attention to the mathematical structure of problems and the context in which problems are put forward, at least not in a systematic way, factors that could be the reason why these problems are so difficult for students. Therefore, in the main, these factors are not taken into account in research results. Nevertheless, for us, they are influential factors on the students' success and on their difficulties in solving problems.

Due to their structure, we already know that we can consider a particular world of problems that is generated by means of a limited set of events and probabilities and the relationships between them (Huerta, 2009). Concretely, with two basic events, 16 probabilities and 18 relationships between these probabilities it is possible to generate this world of problems, which we call the *world of the ternary problems of conditional probability* (Cerdán & Huerta, 2007; Huerta, 2009). Many school problems belong to this world, as do a considerable number of tasks used in research. For this reason their study is relevant. This world of problems has been our subject of research for various years.

Two particular pieces of research will be briefly examined in this paper. Apart from problems and the students' sample, both pieces of work share the same objectives and research methodology. Thus, we will firstly show the shared objectives and methodology and, in continuation, the results we have obtained up to now. Because the research is still not finished and due to the limitation number of pages available for this report, only the global results of the research is shown along with a few comments on them.

### Objectives

Our general objectives in this report are as follows:

1. To identify difficulties of conditional probability problems.
2. To introduce a measure of these difficulties and to carry out these measurements.
3. To consider to what extent the structure and context variables, considered as independent variables, are influential factors on the difficulty of the problems.

### Background: Elements of the framework

Several years ago, we began the study of the conditional probability problems in a systematic and methodical way (Lonjedo & Huerta, 2007; Huerta & Lonjedo, 2006; Cerdán & Huerta, 2007; Carles & Huerta, 2007; Lonjedo, 2007; Huerta, 2009; Edo, 2010). In these works the main object of research is the problems themselves. We identified a particular family of problems that we call *ternary problems of conditional probability* and with the help of a tool we call *trinomial graph* we carried out analytical readings of the problems (Huerta, 2009). An analytical reading of a problem is one that only pays attention to known and unknown data and their relationships but ignoring other elements of the problem that we know are also important. The analytical readings of problems on a trinomial graph allow researchers and teachers to determine the complexity of problems by means of their minimal graph, that is to say, the graph containing the minimum number of required relationships between known and unknown data that solves the problem. At the same time, we use the graph in order to identify problems in which the analytical reading is either arithmetical or algebraical.

Because of the structure of data, ternary problems of conditional probability have been classified into four families and twenty sub-families (Lonjedo, 2007). The L-family of problems is characterized because every problem belonging to it has either 0 or 1 or 2 or 3 known probabilities as a maximum. We call this Level of problems (Huerta, 2009). For each L-family, the C, T-subfamily of problems is characterized by having 0, 1 or 2 known absolute (marginal) probabilities (the Characteristic of a problem) and one unknown probability, which is asked for in problem (the Type of the problem: Type $T_1$ if a conditional probability is asked for, $T_2$ if a marginal and Type $T_3$ if an intersection probability). Therefore, analytically speaking, for each family and sub-family, there is at our disposal a problem representing each one of the class of problems, and consequently its minimal graph. All these previous studies about problems permit us to construct questionnaires of problems to be administered.

All the above-mentioned are elements of the framework in which we place our research on conditional probability problem solving. This framework is still in progress but some of its components can be studied in Huerta (2009).

### *Methodology for assessing difficulties of problems*

Bearing in mind all the above considerations, in order to construct questionnaires for assessing difficulties of problems we first must to decide on what independent variables they might depend. Thus, as task variables (Kulm, 1979) they are, and in order to investigate difficulties depending on structure and context variables, we fixed data format in every problem we put forward. All data referring to probabilities, either known or unknown was expressed in a percentage format, in the case of conditional probability. The reason for taking this decision can be found in various studies (e.g. Watson & Kelly, 2007; Lonjedo & Huerta, 2007). There is a general agreement that students think better using conditional frequencies than with conditional probabilities (see for example, Jones, Langrall & Mooney, 2007).

Thus, for each problem in the questionnaires, we define the following independent variables:

- *Structure variable*, given by the L-family and the C, T sub-family.
- *Context variable*, taking the following values: *Stat-Social*, *Stat-Health* and *Diag-Health*. (See the examples in the annex)

Depending on these two variables, we analyse students' resolutions paying attention to two main dependent variables: process and product variables. Difficulties of problems will be measured throughout the product variable, which takes the following values:

- *Tackled*. Number of students who tackle a problem. We say that a student tackles a problem if we can recognize that student undertakes the problem in some way.
- *Answer*. Number of students that answer the question posed in the problem. The answer to the question may be anything that students declare is the answer to the question posed.
- *Number*. Number of students that respond to the question of the problem by means of a correct number (a percentage in this case).
- *Description*. Number of students that attach an expression to the numerical answer, describing what this number is measuring. For a given number, the attached expression may be correct or incorrect whether the number is correct or not.

By means of these variables we define the following difficulties of problems:

- *Appreciated difficulty (ADP)* of a problem. This tries to measure students' difficulties before the process of solving the problem starts. Therefore, we are supposing that students, after reading the problem, decide whether to tackle the problem or not. If they do not, we suppose it is because he/she appreciates that the problem is too difficult to be solved.
- *Problem's difficulty (PD)*. This difficulty will inform us to what extent it is difficult to give an answer to the question in problem.

- *Problem solution difficult (PSD).* In relation to the previous difficulty, this one indicates to us how difficult is to give a correct answer to the problem.
- *Difficulty of Correct Description of the solution of the problem (CorrectDescD).* In this case, we will obtain information about students' difficulties in giving a correct description of the event which is measured by the given number as a solution.

Differences between difficulties may be also appreciated through their measures and the way we measure them. Indeed, each one of these difficulties is measured in percentages, as a result of applying the following formulae:

$$ADP = 100 - \left(\frac{tackled}{number\ of\ students}\right)\times 100 \quad PD = 100 - \left(\frac{answer}{tackled}\right)\times 100 \quad PSD = 100 - \left(\frac{number}{tackled}\right)\times 100$$

$$DescD = 100 - \left(\frac{Description}{answers}\right)x100 \quad CorrectDescD = 100 - \left(\frac{CorrectDesc}{Description}\right)\times 100$$

These difficulties are ranked in an [0, 100]-interval. If a difficulty is measured by one 0, then this means that the problem does not pose that particular difficulty for any student whereas, on the opposite side of the interval, a difficulty which is measured by one 100 is present in the problem for all students.

***Results from two independent, but related, researches***

1. Research on difficulties of problems from $L_0$-family in students15-16 aged without previous instruction on conditional probability.

The $L_0$-family of problems can be divided into three sub-families (Huerta, 2009; Edo, 2010): $C_0T_1$, $C_1T_1$ and $C_2T_1$. There are 6 *basic-options* of problems representing to this family: one from $C_0T_1$; two from $C_1T_1$; and finally three from $C_2T_1$. Accordingly, we constructed one basic questionnaire containing 6 problems. Owing to the fact we finally consider three contexts, in order to explore difficulties of problems depending on contexts, we prepared a total of 18 problems (some examples are given in the annex). All those problems were distributed into 6 questionnaires each containing 3 + 3 problems structurally isomorphic in pairs, but put forward in different contexts. For each questionnaire, and from students' resolution of problems, the possible influence of the context in the difficulties of problems becomes observable for researchers. Furthermore, at the same time, the influence of the structure of problems on these difficulties is also observable. The next table (Table 1) shows us the properties of each one of the questionnaires that were finally administered to students.

| Questionnaire | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 |
|---|---|---|---|---|---|---|
| Contexts | All | All | All | All | All | All |
| Sub-family (All $T_1$) | $C_0$; $C_1$ | $C_0$; $C_2$ | $C_1$; $C_2$ | $C_1$; $C_2$ | $C_1$; $C_2$ | $C_2$ |

**Table 1: Questionnaires assessing difficulties of problems from $L_0$-family.**

Questionnaires were administered to 165 students aged 15-16 years. Students solved problems during their mathematics class. They received no teaching on conditional probability. Table 2 shows the distribution of the number of students in each questionnaire.

| Questionnaires | $Q_1$ | $Q_2$ | $Q_3$ | $Q_4$ | $Q_5$ | $Q_6$ |
|---|---|---|---|---|---|---|
| Number of students (165) | 45 | 33 | 21 | 19 | 23 | 24 |

**Table 2: Number of students that solved each questionnaire**

The 990 students' resolutions provided the following global results (Table 3):

| $L_0$-Problems' resolutions | Tackled | Answer | Number | Descriptions | Correct Desc |
|---|---|---|---|---|---|
| 990 | 712 | 585 | 214 | 320 | 139 |
| Difficulties<br>(%) | ADP<br>(28.1) | PD<br>(17.8) | PSD<br>(**69.9**) | DescD<br>(45.3) | CorrectDescD<br>(**56.6**) |

**Table 3: Global results en frequencies and difficulties of problems from $L_0$-family, in percentages.**

If we take into account contexts and structures, the following tables (Tables 4 & 5) show us to what extent these are influential factors:

| Structure/ Difficulty | ADP | PD | PSD | CorrectDesD |
|---|---|---|---|---|
| $C_0T_1$ | 8.5 | 13.5 | 59.5 | **69** |
| $C_1T_1$ | 27.2 | 17.8 | **64.9** | 49.6 |
| $C_2T_1$ | **41.9** | 21.4 | **65.5** | 51.4 |

**Table 4: Difficulties of problems from $L_0$-family, depending on their structure (%)**

| Context/difficulty | ADP | PD | PSD | CorrectDesD |
|---|---|---|---|---|
| Stat-Social | 20.9 | 13.4 | 50 | 41.2 |
| Stat-Health | 23 | 18.1 | **75** | **75.7** |
| Diag-Health | **40.3** | 23.3 | 62.9 | 55.1 |

**Table 5: Difficulties of problems from $L_0$-family, depending on contexts (%)**

2. Research on difficulties of ternary problems (problems belonging to all families) in student teachers of high school mathematics.

In this research we studied the difficulties of ternary problems in a sample of 54 students doing a Professional Master in Secondary Mathematics Teaching at the Universitat of València (Spain). Not all of them were Mathematics graduates, as there were also graduates in Economics, Engineering and Architecture. However all of them had taken courses in probability during their regular studies.

In order to construct a questionnaire for assessing difficulties of problems, and given that we did not have prior experience on problem resolutions in students belonging to this sample, many previous decisions had to be taken into account. Only the experience with $L_0$-problems could help us in doing this. In addition to this, previous to deciding what problems were going to be included in the questionnaire, we carried out a theoretical study of problems at every level, in order to determine *basic-options* of problems representing all families and sub-families, as had been done for the $L_0$-family. Based on this study, among others, the following decisions were taken:

- The questionnaire consists of 7 ternary problems distributed as follows: One problem from $L_0$-familiy, two problems from $L_1$-family, three problems from $L_2$-family, and one problem from $L_3$-family.
- In every problem, the data format is percentages.
- *Stat-social*, *Stat-Health* and *Diag-Health/man* are the contexts in which problems are put forward. Differences between Diagnostic Test in Health or Manufacture must be found in the elements of the context: people or manufactured pieces that are tested.
- If a problem has a complex analytical reading, either arithmetical or algebraical, we decided to avoid extra difficulties by contextualizing the problem in a context that we know is not very influential in the difficulties. This is the case, for example, in the problem that belongs to the $L_3$-family. We already know that all problems from this family have an algebraical reading. On the other hand, from the previous research, we also know that Stat-social is a less influential context on the difficulties than the other two we considered. Therefore, we decided to formulate the problem from $L_3$-family in a Stat-social context. Conversely, the problem from $L_0$-family is formulated in the most influential context, the Stat-Health context.

Finally, these are the characteristics of the questionnaire that was administered to the students (Table 6):

| Problem | P1 | P2 | P3 | P4 | P5 | P6 | P7 |
|---|---|---|---|---|---|---|---|
| Structure | $L_0C_2T_1$ | $L_2C_1T_1$ | $L_3C_0T_2$ | $L_1C_0T_1$ | $L_1C_1T_1$ | $L_2C_0T_1$ | $L_2C_1T_1$ |
| Context | Stat-Health | Diag-Manufac. | Stat-Social | Stat-Social | Diag-Health | Stat-Social | Stat-Health |
| Analitical readings | Arith. | Arith. | Algeb. | Arith. | Arith. | Algeb. | Arith. |

**Table 6: Questionnaires assessing difficulties of ternary problems**

The 378 students' resolutions produced the following global results (Table 7):

| Ternary Problems' resolutions | Tackled | Answer | Number | Descriptions | Correct Desc |
|---|---|---|---|---|---|
| 378 | 369 | 269 | 109 | 183 | 138 |
| Difficulties (%) | ADP (2.4) | PD (27.1) | PSD (**70.5**) | DescD (32) | CorrectDescD (24.6) |

**Table 7: Global results en frequencies and global difficulties of problems in %**

In a similar way, as was done in the above example, the following table (Table 8) provides us with information concerning the difficulties of the problems depending on the family to which they belong:

| Level | ADP | PD | PSD | DescD | CorrectDesD |
|---|---:|---:|---:|---:|---:|
| $L_0$ | 1.8 | 16.7 | 68.5 | 28.9 | 9.4 |
| $L_1$ | 1 | 26.2 | 68.2 | 29.1 | 25 |
| $L_2$ | 4.3 | 28.4 | 69.8 | **38.7** | 27.9 |
| $L_3$ | 0 | **37** | **79.6** | 20.6 | 11.1 |

**Table 8: Difficulties of problems depending the family they belong to, in (%)**

And, finally, the following table shows the difficulties of problems depending, this time, on the contexts in which the problem was presented:

| Context | ADP | PD | PSD | DescD | CorrectDesD |
|---|---:|---:|---:|---:|---:|
| Stat-Social | 1.8 | 35.2 | 71.7 | 28.1 | 14.9 |
| Stat-Health | 4.6 | 31.1 | **74.6** | 33.8 | 27.7 |
| Diag-Health/Man. | 1 | 11.2 | 64.5 | 34.7 | 33.9 |

**Table 9: Difficulties of problems depending on contexts, in (%)**

### *Discussion and conclusions*

From a global point of view, we can conclude that ternary problems of conditional probability are difficult for students. Finding a correct answer to a problem, which is basically finding a conditional probability when three known data are given in a context, is a difficult task for every student, whether they are from secondary school or graduated. But, in general, at the beginning of the process, it seems that these difficulties are not especially appreciated by students (ADP between 28%, in the first study, and 2.4% in the second one) when they tackle the problems and try to give an answer to the question posed in them (PD between 17.8% and 27.1%). However, the answers given are usually incorrect in a very high percentage of cases (PSD of 69.9% in the first study, 70.5% in second one). Among those students who give an answer, some of them do not describe the number, given as the answer, by means of an expression representing the event and the number is its measure (45.5% in first, 32% in second). That is to say, for these percentages of students, the answer is simply a number, neglecting what this number is measuring and why. Among the students that describe the number with an expression, there is a percentage of students that do it incorrectly (56.6% in the first one, 24.6% in the second one). In some of these incorrect expressions it is usual to recognize one of the most common misconception in conditional probability: the conditional probability by intersection probability

(Lonjedo & Huerta, 2007; Lonjedo, 2007), even if the given number is a correct number and the students are either at secondary school or graduated.

Structures and contexts appear as influential factors on the difficulty of problems, as we have shown above. Thus, for example, in the second study, we can see how the difficulty of giving an answer to problems increases as the number of known conditional probabilities in text of problems also becomes greater. However, when an answer is given, the difficulty of giving the correct answer is not as sensitive to this factor, except in the $L_3$-problem. It seems that there are other structure-based factors that are influential on this difficulty, or on the combination of structure+context factors, as we can see in the next table (Table 10).

| Structure | Context | ADP | PD | PSD | PDescD | PCorrectDesD |
|---|---|---|---|---|---|---|
| $L_0C_2T_1$ | Stat-Health | 1.8 | 16.7 | 68.5 | 28.9 | 9.4 |
| $L_1C_0T_1$ | Stat-Social | 0 | 35.2 | 63 | 22.8 | 22.2 |
| $L_1C_1T_1$ | Diag-Health | 1.8 | 17 | 73.6 | 34.1 | 27.6 |
| $L_2C_0T_1$ | Stat-Social | 5.5 | 33.3 | 72.5 | 41.2 | 10 |
| $L_2C_1T_1$ | Stat-Health | 7.4 | 48 | 82 | 42.3 | 26.7 |
| $L_2C_1T_1$ | Diag-Manu | 0 | 5.5 | 55.5 | 35.3 | 39.4 |
| $L_3C_0T_2$ | Stat-Social | 0 | 37 | 79.6 | 20.6 | 11.1 |
| Test | — | 2.4 | 28.8 | 70.5 | 32 | 24.6 |

**Table 10: Difficulties of problems depending on contexts+structure (%)**

Finally it is reasonable to think that contexts are an influential factor on the difficulties of problems. But difficulties are not sensitive to these influences in the same measure. Thus, while the Diag-Health context is an influential factor on the appreciated difficulty of the problem for secondary school students, this is not the case for future teachers. This could mean that the context is not a well-known context for the former but that it is for the latter. On the contrary, if we pay attention to the problem solution difficulty, it is common to both samples that the Stat-Health context is the most influential. Considered in an isolated way or in combination with structure, difficulties are very sensitive to it. This context is characterized because it belongs to a more general situation called *Les situations causalistes* (Henry, 2005), that particularly in the Health context is very influential: every time someone treats himself/herself with a medicine, then he/she is cured of the illness, but if this is not so then he/she does not. Obviously, there are no conditional probabilities involved.

Although the results we just show in this report are global and their analysis is not yet finished, some suggestions can be made for the future, both for researchers in probabilistic education and for mathematics teachers. Researchers could take into account that tasks used to investigate students' behaviour have different difficulties depending on the structure of their data and the context chosen for putting forward the task. It is a hypothesis to suppose that varying structure and context in tasks imply that students' behaviour may also be different in each case, or not. This hypothesis might be contrasted.

On the other hand, teachers should be aware that there are families of conditional probability problems that contain problems with different difficulty degrees, again depending on structures of data and context. Perhaps a teaching model could be considered which was based on solving problems belonging to these families, organized sequentially, and exploring through them as contexts as possible. This could potentially be a good way to improve students' competence in solving conditional probability problems.

## *References*

Carles, M.; Huerta, M. P. (2007). Conditional probability problems and contexts. The diagnostic test context. In D. Pitta-Pantazi & G. Philippou (Eds.) *Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education, CERME 5,* 702-710.

Cerdán, F.; Huerta, M. P. (2007). Problemas ternarios de probabilidad condicional y grafos trinomiales. *Educación Matemática*, 19 (1), 27-62.

Díaz, C.; De la Fuente, I. (2007). Assessing students' difficulties with conditional probability and Bayesian reasoning. *International Electronic Journal of Mathematics Education*, 2 (3), 128 -143.

Edo, P. (2010). *Estudios sobre la familia de problemas de probabilidad condicional de nivel $N_0$* (Studies about the family of conditional probability problems of level $L_0$). Unpublished Memory of degree in Master of Research in Didactic of Mathematics. University of Valencia.

Estrada, A; Díaz, C. (2006). Computing probabilities from two-way tables: an exploratory study with future teachers. *Proc. ICOTS-7*. Retrieved June 7, 2009, from http://www.stat.auckland.ac.nz/~iase/publications/17/C413.pdf.

Henry, M. (2005). Modélitation en Probabilités conditionnelles. In M. Henry (Ed.) *Autour de la modelisation en probabilités*. IREM de Franche-Comté, (6) 173 - 185.

Huerta, M. P. (2009). On Conditional Probability Problem Solving Research —Structures and Context. In M. Borovcnik & R. Kapadia (2009), Special issue on "Research and Developments in Probability Education". *International Electronic Journal of Mathematics Education*, 4 (3), 163-194.

Huerta, M. P.; Lonjedo, M. A. (2006). The nature of the quantities of in a conditional probability problem solving behaviour. In M. Bosch (Ed.) *European Research in Mathematics Education IV. Proceedings of the Fourth Congress of the European Society in Mathematics Education, CERME 4*. Sant Feliu de Guíxols (Spain).

Jones, G. A.; Langrall, C. W.; Mooney, E. S. (2007). Research in Probability (Responding to Classroom Realities). In F. Lester (Ed.), *Second Handbook of Research on mathematic teaching and learning*, 909-956. NCTM.

Kulm, G. (1979). The classification of Problem-Solving Research Variables. In G. A. Golding & C. E. McClintock (Ed.), *Task Variables in Mathematical Problem Solving*, 1-22. ERIC

Lonjedo, M. A.; Huerta, M. P. (2007). The same problem in three presentation formats. Different percentages of success and thinking processes. In D. Pitta-Pantazi & G. Philippou (Eds.) *Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education, CERME 5,* 732-741.

Maury, S. (1984). La quantification des probabilités : analyse des arguments utilises par les élèves de classe de seconde, *Recherches en Didactique des Mathématiques,* 5, (2), 187-214.

Ojeda, A. M. (1996). Contextos, representaciones y la idea de probabilidad condicional. In F. Hitt (Ed.), *Investigaciones en Matemática Educativa*. México: Grupo Editorial Iberoamericana, 291-310.

Watson, J. M.; Kelly, B. A. (2007). The development of conditional probability reasoning. *International Journal of Mathematical Education in Science and Technology*, 38:2, 213-235.

## Annex: Example of problems used in the researches.

Problem in Level $L_0$ (15-16 years old students. Stat-social context)

The 4th grade class is made up of 30 students between boys and girls. Among the students, there are 7 boys who wear glasses, 10 girls who do not use them, and 8 boys who do not wear them. Among boys in the class, what percentage wears glasses?

Problem in Level $L_1$ (graduated students. Diag-Health context)

A population at risk for tuberculosis is subject to the tuberculin test. Different studies show that 57% of the population suffers from tuberculosis, and among those with tuberculosis 59.6% give positive in test. In addition, it is known that 13% did not have tuberculosis but is positive in the test. Among those who are positive in the test, what percentage is suffering from tuberculosis?

Problem in Level $L_2$ (graduated students. Stat-Health context)

A population suffers from eye infection. Of these, 42% are treated with a new antibiotic. The results show that, of those treated with the antibiotic, 83.3% were cured, and people who have not cured 14.9% were treated with the antibiotic. Among those who have been cured, what percentage has not been treated with the new antibiotic?

Problem in Level $L_3$ (graduated students. Stat-social context)

Of the girls in a high school, 37.5% wear glasses. Of the boys, 28.6% wear glasses. Of those who do not wear glasses, 50% were boys. Among high school students, what percentage are girls?

[1] EDU/2008-03140/EDU Project, funded by Ministerio de Ciencia e Innovación, España.