Cluster and Factor Analysis of Students' Responses to the CPR Questionnaire

Carmen Díaz
University of Huelva, Spain, Faculty of Psychology

Helena Bacelar-Nicolau
University of Lisbon, Portugal, Faculty of Psychology

Carmen Batanero
University of Granada, Spain, Faculty of Education
batanero@ugr.es

1. Introduction

Conditional probabilities and the Bayes theorem are highly relevant in the field of psychology, where these types of reasoning appear in diagnosis, evaluation, decision-making and application of statistical inference in experimental research. However, recent research related to conditional probability and Bayesian reasoning suggests the existence of different psychological biases in reasoning about conditional probability. These difficulties include confusing conditional and causal reasoning (Falk, 1989), the fallacy of the time axis, or belief that in P(A/B), event B should always precede in time event A (Gras & Totohasina, 1995; Ojeda, 1996), problems in defining the conditioning event and misunderstanding of independence (Sánchez, 1996; Truran & Truran, 1997). As regards Bayesian reasoning (see a summary in Koehler, 1996), early research by Tversky and Kahneman (1982) suggested that people do not employ this reasoning intuitively and established the robustness and spread of the base-rate fallacy in students and professionals (Bar-Hillel, 1997). While this previous research concentrated on isolated points of understanding, at the University of Granada we developed a comprehensive instrument (CPR questionnaire, Conditional Probability Reasoning) that can be used to assess in a reasonable time and in just one application the different biases and misunderstanding of these concepts in the same student (Díaz, 2004). In this paper we use cluster and factor analysis to explain that the biases described above do not appear to be related to formal understanding and competence in solving conditional probability problems but rather are independent to one another. We finally conclude with some recommendation to improve the teaching of conditional probability to psychologists.

2. Method

Data are taken from 590 psychology students in 4 different Spanish Universities who completed the CPR questionnaire with a total of 18 items. Some items had several parts (e.g., item 6 and 17), which are scored independently. The building of the questionnaires was based in a rigorous methodological process, including expert judgment, item trialling, validity and reliability assessment and Bayesian estimation of items' psychometric features. All the students in the different samples in this study were in the first year of Psychology and all of them followed an introductory statistics course with the same programme. The different tests were all given after the students had been taught conditional probability and Bayes theorem. The questionnaires were completed as an activity in the course of data analysis, and the students were asked to study the topic in advance. Our first approach to the reliability of the instrument was carried out by computing the coefficient Cronbach's Alpha, that gave a moderate value (Alpha=0.721). A second estimation of reliability used test-retest in a sample of 106 students that completed the questionnaire twice (with a month between testing times) and provided a reliability coefficient of 0.859, which is reasonably high. A theoretical analysis of the questionnaire content as well as the results from experts' judgement served to justify content validity comparing the content evaluated by each item to the semantic units included in the semantic definition.

We analysed the structure of responses to the questionnaire and compared with the assumed structure of the construct (Muñiz, 1994; Martínez Arias, 1995). We performed both an exploratory factor analysis and a hierarchical cluster analysis. The factor extraction method was principal components, which is the most conservative method, as it does not distort the data structure and we used the Varimax rotation (orthogonal rotation; maximizing variance of the original variable space). Cluster analysis hierarchical model was based on complete linkage aggregation criterion and Ochiai coefficient applied to the data recoded into a binary scale (correct-incorrect responses to each item by the different students) (Bacelar-Nicolau, & Figueira, 1989; Bacelar-Nicolau & Nicolau, 1990; Bacelar-Nicolau, 2003). We expected the analysis confirm a main underlying construct, and at the same time we also expected to find other factors or features that include the biases described in the literature and that would not correlate with the mathematical problem solving competence of students.

3. Results and discussion

Results showed the following percentages of students with problems in understanding conditional probability: not correctly restricting the sample space in conditional probability (26.3%), base-rate fallacy (16.3%), incorrect identification of Bayes' formula (29.1%), confusion of independence with mutual exclusiveness (30.1%), only considering independence in diachronic experiments (18.4%), difficulties in reading conditional or joint probabilities from a 2x2 table (between 30 and 40% in the different parts of Item 6), confusing a conditional probability with its inverse (33%), conjunction fallacy (7%), difficulties in computing probabilities when the time axis is inverted (80%). To explore our conjecture that biases on conditional probability reasoning are unrelated to mathematical performance in the tasks, factor analysis was first carried out on the set of responses to the items using the SPSS software. Table 1 shows the factor loadings (correlations) of items with the different factors after rotation.

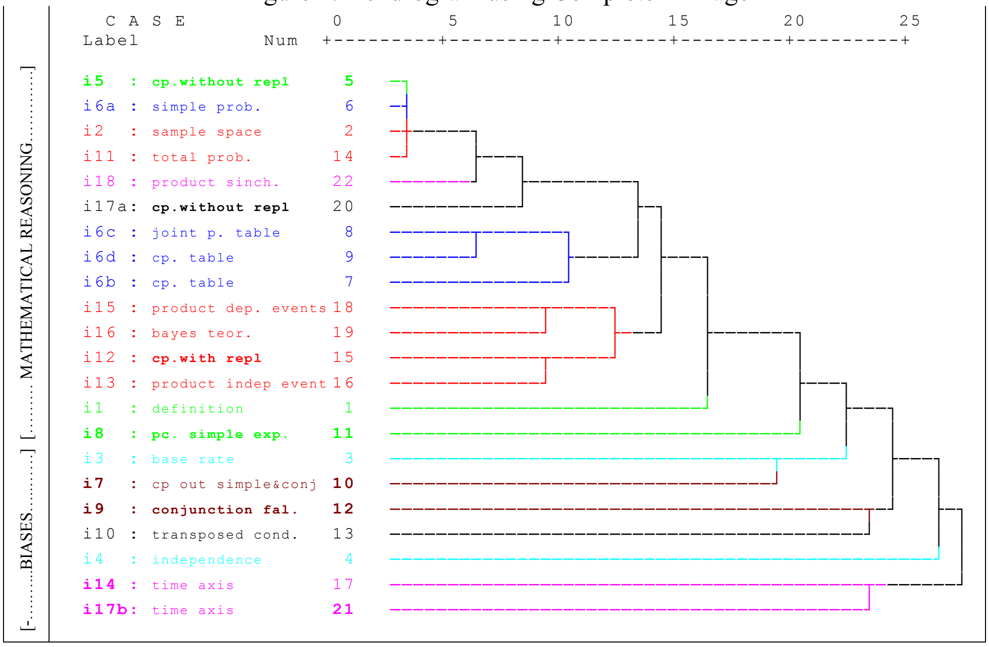
Table 1. Factor Loadings for Rotated Components in Exploratory Factor Analysis of Responses to Items

Item	Component						
	1	2	3	4	5	6	7
i16. Bayes rule	**.76**						
i11. Total probability	**.76**						
i15. Product rule in dependent events	**.75**						
i13. Product rule in independent events	**.67**						
i12. Conditional probability with replacement	**.43**		**.42**				
i6b. Conditional probability. Table		**.79**					
i6c. Joint probability. Table		**.77**					
i6a. Simple probability. Table	**.32**	**.61**					
i6d. Conditional probability. Table		**.61**					
i8. Conditional probability in single experiment			**.67**				
i1. Definition			**.59**				
i2. Sample space	**.40**		**.45**				
i17b. Time axis fallacy				**.71**			
i14. Time axis fallacy				**.70**			
i7. Cond prob. from joint and compound probability					**.66**		
i9. Conjunction fallacy					**.62**		
i5. Conditional probability, without replacement			**.39**		**.44**		
i17a. Conditional probability, without replacement						**.66**	
i10. Transposed conditional /causal-diagnostic						**-.65**	
i4. Independence /mutually exclusiveness							**.68**
i3. Base rates/ Bayes rule	**.34**						**.48**
i18. Product rule dependent events, diachronic				**.35**			**-.46**

Seven factors with eigenvalues higher than 1 were found that explained the following percentages of the total variance: 21% (first factor), 7 % (second factor), and about 6% in the remaining factors; that is, a total of 59% of the variance was explained by this set of factors, which suggests the specificity of each item, and the multidimensional character of the construct, even when there is a common part shared by all of the items. These percentages of variance also revealed the greater importance of the first factor, to which most of the resolutive problems contribute; in particular solving Bayes' problems had the higher contribution, followed by solving total probability and compound probability problems. All of these problems require a solving process with at least two stages, in the first of which a conditional probability is computed, which is used in subsequent steps (e.g. product rule). We could interpret this factor as solving complex conditional probability problems ability.

Computing simple, joint and conditional probability from a two-way table (item 6) appeared as a separate component, probably because the task format affected performance, a fact which has also been noticed by Ojeda (1996) and Gigerenzer (1994), among other researchers. A third factor showed the relationships between definition, sample space and computation of conditional probabilities in, with and without replacement situations; we interpreted this factor as Level 4 reasoning in Tarr and Jones (1997) classification. The remaining factors suggested that the different biases affecting conditional probability reasoning that are described in the justification, appeared unrelated to mathematical performance in problem solving, understanding, building the sample space and computing conditional probability, and to Tarr and Jones's (1997) level 4 reasoning (as related items were not included in the three first factors). Each of the biases (transposed conditional, time axis fallacy, conjunction fallacy, independence/mutually exclusiveness/synchronic setting) also appeared unrelated to one another; in some cases some of them were opposed or related to some semantic units in the mathematical component of understanding conditional probability. For example, independence was linked to the base rate fallacy (where people have to judge whether if the events are independent or not) and opposed to the idea of dependence.

Figure 1. Dendrogram using Complete Linkage



The previous results were confirmed and/or complemented by cluster analysis hierarchical model. The dendrogram (and associated level of coefficient values) shows that only one part of the items clearly split in three clusters. The first cluster includes items 5, 6a, 2, 11, 18 and 17a. This cluster could be described as a "conditional probability general reasoning" where items are significantly correlated with each one of the different principal components in the factor analysis. The second cluster corresponds to the items 6b, 6c and 6c so it gives better explanation to the meaning of the second principal component in the factor analysis. The third cluster includes all the items high correlated with the first component that are 15, 16, 12 and 13, thus we interpret it as solving complex conditional probability problems ability. These three clusters merge together into a global cluster, which may represent the mathematical skills involved in the conditional probability reasoning. Then the remaining items join the global cluster in an almost perfect chain merging way that confirms that each of the psychological biases assessed in the CPR questionnaire are independent to the mathematical reasoning. Differences in the two analyses may be explained by the fact that Table 1 shows the results after rotation and small correlations (under .30) are deleted. Items are also re-ordered according their contribution to the first factor and some items contribute to more than a factor. In summary, these results supported our previous hypotheses that biases in reasoning about conditional probability are unrelated to mathematical performance in problem solving and, at the same time, support construct validity evidence for the questionnaire. Moreover it provides information about potential biases students might hold that were used in the design of the teaching experience in the next step of this research.

ACKNOWLEDGEMENT: This work was supported by the project SEJ2004-00789 and was developed on the scope of the cooperation project between the University of Granada and the Laboratory of Statistics and Data Analysis of the FPCE, University of Lisbon.

References

Bacelar-Nicolau, H. & Figueira M.L. (1989) Análise exploratória da evolução do auto-conceito nos adolescentes através de modelos probabilísticos de classificação hierárquica. *Psiquiatria clínica*, vol. 10, nº 1, 43-48.

Bacelar-Nicolau H. & Nicolau F.C. (1990). Analyse classificatoire de l'identité nationale chez les portugais. Une étude exploratoire fondée sur le coefficient d'affinité. *Mathématiques, Informatique et Sciences Humaines*, 28eme année, nº 109, 55-64.

Bacelar-Nicolau, H. (2003). *Introdução à Análise Classificatória Hierárquica Ascendente: Modelos de ACHA*. Notas e Comunicações do LEAD

Bar-Hillel, M. (1987). The base rate fallacy controversy. In R. W. Scholz (Ed.), *Decision making under uncertainty.* (pp 39 – 61) Amsterdam: North Holland.

Díaz, C. & de la Fuente, I. (2006). Assessing psychology students' difficulties with conditional probability and bayesian reasoning. In A. Rossman y B. Chance (Editores). *Proceedings of ICOTS-7.* Salvador (Bahia): International Association for Statistical Education. CD ROM

Gras, R & Totohasina, A. (1995). Chronologie et causalité, conceptions sources d'obstacles épistémologiques à la notion de probabilité. *Recherches en Didactique des Mathématiques*. 15(1), 49 – 95.

Falk, R. (1989). Inference under uncertainty via conditional probabilities. In R. Morris (Ed.), *Studies in mathematics education: Vol.7. The teaching of statistics* (pp. 175-184). Paris: UNESCO,

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

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavior and Brain Sciences, 19*, 1-54.

Ojeda, A. M. (1996). Contextos, representaciones y la idea de probabilidad condicional (Contexts, representations and conditional probability). In F. Hitt (Ed.), *Investigaciones en matemáticas educativas* (pp.291-310). México: Grupo Editorial Iberoamericano.

Sánchez, E. (1996). Dificultades en la comprensión del concepto de eventos independientes (Difficulties in understanding independent events). In F. Hitt (Ed.), *Investigaciones en Matemática Educativa*, pp. 389–404. México: Grupo Editorial Iberoamericano.

Tversky, A, & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153-160).. New York: Cambridge University Press.