Regression trees and decision rules applied to the study of perplexity: Issues and Methods

Inés Mª Gómez-Chacón, J. Tinguaro Rodríguez

Universidad Complutense de Madrid, igomezchacon@mat.ucm.es,

This paper explores the use of the CRT (Classification and Regression Tree) methodology to analyse data from a fuzzy rating scale-based questionnaire. Based upon a questionnaire to assess the state of perplexity in mathematics undergraduate students, the rule structure obtained from the CRT analysis is reported. We anticipate these findings may be of interest both to evaluate the interplay between cognition and affect as well as to researchers in the Fuzzy Logic field.

Keywords: Emotions, Heuristics, Fuzzy set, Cognition and Emotion, Mathematics, Regression trees

Introduction

In this paper, we will focus on perplexity in mathematics. In the studies concerning determination of affective pathways during the solving of problems (Goldin 2000; Gómez-Chacón, 2000, in press), the state of perplexity or puzzlement is considered to be one of the interesting emotional states into which the individual can drift along positive or negative pathways when solving the problem. Perplexity does not in itself have unpleasant overtones—but bewilderment, can include disorientation and a sense of having "lost the thread". If problem solving continues, a lack of perceived progress may generate frustration, where the negative affect becomes more powerful and more intrusive, but where there is still the possibility that a new approach will move the solver back to the sequence of a predominantly positive affect. The studies mentioned above show the need to understand and to know, in depth, the benefits that this state can achieve for the teaching and learning of mathematics.

A big challenge today is to improve the methodological tools for evaluating affect detection systems. This need is motivating studies in trying to explain this gradualness in the processing of affective mechanisms. Further research is necessary to reference the identification, discrimination, and the unclear boundary between the cognitive and affective processes. This paper explores the use of decision trees to analyse data from a fuzzy rating scale-based questionnaire. Thus, here we will report our results in the form of a tree structure providing rules to assess the state of perplexity in mathematics undergraduate students, built upon the basis of the previous questionnaire.

The present research is primarily exploratory for two reasons: 1) perplexity has been scantly analyzed in mathematics and educational psychology; and 2) the use of the CRT methodology to analyse data from a fuzzy rating scale-based questionnaire is a new development. The theoretical background and empirical studies related to perplexity need to be developed.

Theorists of science and mathematics (Lakatos 1976) claim that mathematical reasoning and complex problem solving are typical cognitive tasks in which perplexity is directly involved. However, an analysis of the psychological (cognitive and affective) processes involved in it lacking, in order to clarify the definition. (.Studies about confusion (Silva 2010) indicate an appraisal structure of this emotion of novelty-complexity that is reflected in a state of uncertainty and

comprehensibility, reflecting an inability to understand. Smith and Ellsworth's (1985) appraisal model maintains that in order to differentiate emotional experiences some dimensions are key (pleasantness, attentional activity, control, certainty, goal-path obstacle and anticipated effort). Taken together in this study, these dimensions and mathematical cognitive processes could provide information about perplexity, a knowledge emotion that is rarely studied, and illustrate the relationships between these cognitive variables and emotion in order to deepen understanding of its nature.

The following hypotheses guided the work: (1) the emotions associated with the perplexity state could be positive, negative, or neutral; (2) those participants with more positive ability to cope with the situation (control), better ability to predict, or with a wider understanding, are those who will better handle their state of perplexity in the reasoning.

Fuzzy rating method for questionnaires

The method of fuzzy rating scale applied in this research was introduced by Hesketh, Pryor, and Hesketh (1988) and subsequently developed in various studies by Gil et. al., (2015). The fuzzy approach is based on the idea that, in some cases, it is not reasonable to say that an object has to either verify a property or not verify it (Zadeh, 1975). Objects or people may exhibit some properties only partially—i.e. up to a certain extent or degree. In many of the conducted researches the evaluation of the emotion parameters is qualitatively performed through reports, interviews, recording observations or, if it is quantitative, through Likert scales. In the case of Likert scales (based on an implicit subjacent numerical continuum), such kind of imprecise information is lost, since finally just a single category has to be chosen (which excludes the representation of a potential hesitation between categories).

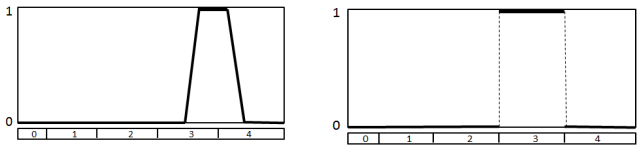

Fig. 1. Examples of Fuzzy sets valuating emotion

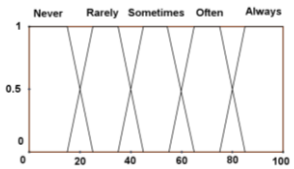

Fig. 2. Fuzzy sets modelling linguistic variable

In this sense, fuzzy logic allows relaxing this constraint by admitting valuations to be given in the form of fuzzy numbers over the subjacent numerical scale. That is, in this setting each possible numerical evaluation is assigned a degree of membership or verification, between 0 and 1, representing the validity of such number as a measure of the observed emotional phenomenon (Fig.1).

The idea of using these fuzzy sets to describe imprecise terms is closely linked to the concept of linguistic variable introduced by Zadeh (see Zadeh, 1975). A linguistic variable is considered one that takes linguistic values, which is less accurate than the use of numbers. For example, a linguistic variable used to evaluate 'confidence' may take the (linguistic) values: never, rarely, sometimes, often and always. Each of the linguistic terms that can take the variable is modeled by a fuzzy set (see Fig. 2, and notice the subjacent numerical scale that accompanies the linguistic descriptions). There are values of the variable that can be assigned up to a degree to two of these fuzzy sets (e.g.

through disjunction) and, therefore, the boundaries between two consecutive linguistic values can be made flexible.

In this study, trapezoidal fuzzy sets (or fuzzy numbers) were used to perform evaluation on a continuum, assigning a membership degree between 0 and 1 to each point of the interval [0,100] (see Fig. 2 again). Notice that trapezoidal numbers allow representing a continuum of prototypes (i.e. elements that are assigned membership degree 1) with a linear decay.

Regression trees

The CRT methodology is a data mining approach widely employed to develop 'IF-THEN' rule models in order to explain the behaviour of a variable of interest (the dependent variable) in terms of logical conditions over a set of explanatory or independent variables (see Breiman et al., 1984). As such, classification and regression trees have been successfully applied to different data-analysis tasks, such as segmentation, stratification, forecasting, data reduction, variable selection, etc., in wide variety of practical contexts (see Strobl et al., 2009).

Particularly, the CRT methodology allows determining a subset of the available independent variables as well as a set of conditions over these variable's values that separate the data into groups as homogeneous as possible in terms of the values of the response or dependent variable. To this aim, the CRT method performs successive dichotomous splits of the data by identifying both the independent variable and its cut-point that provide the greatest variability (i.e. variance) reduction at the split data groups verifying either condition (i.e. being greater or lower than such cut-point). This process starts at the root node containing all the available data, and is iterated at the resulting nodes or groups until some stopping criterion (usually concerning the depth or number of splits or the sample size in the undivided nodes) is reached. The nodes that are left undivided at the end of this process, usually known as leaves, provide conditional response-variable distributions (assumed to be as homogeneous as possible) given the conditions or premises formed by the conjunction of the different branches (i.e. splits) that separate each leaf from the root node.

Notice that CRT is a data analysis methodology with almost no assumptions, and particularly that it is a non-parametric and distribution-free model-building method (e.g. no normality or independence assumptions are made). For these reasons, CRT is especially useful as an exploratory tool allowing to uncover some relationships and patterns in the data that may be expressed in logical form.

In this work we apply this regression tree methodology to develop a rule model capturing the relationships between a numerical dependent variable, measuring either the intensity of perplexity or pleasure experienced by students while solving a mathematical problem, and a set of independent variables measuring the intensity of other emotions that may appear in consonance with perplexity. Our aim, at least in a first stage, is basically exploratory; that is, we do not pursue a complex mathematical model of those relationships of perplexity with other emotions, but rather a simple model describing the most significant relationships in terms that may be checked intuitively. In this sense, we found that the CRT methodology fitted this aim quite well.

Research questions and methodology

Research Questions

We particularly pursued the following research questions: Research question 1: What emotions and cognitive appraisal processes have more influence on the state of perplexity? Research questions 2: How pleasant is being in the state of perplexity? What variables are related to the dimension of pleasantness?

Participants and instrument

Data was collected in 2014 from 100 (56 women and 44 men, aged between 22 and 23) Caucasian undergraduates working toward a BSc. in mathematics. All of the participants were in their last year of academic study, and were distributed into three training groups established by the academic institution. They were following advanced courses in several areas of geometry, algebra, probability and analysis. With regard to solving problems, the students had been introduced to the problem solving heuristics and they received training as students and in one subject related to advanced professional knowledge, practice and relationship skills relevant to teaching. They had not received any special training about backtracking heuristics.

The work dynamic was individual work and began with a paper and pencil resolution of four tasks (problems), each of one hour and a half duration. One example problem, the results of which will be analysed later in section of results, is shown below:

> Paths: How many paths consistent in a series of horizontal segments and / or vertical can be counted in the figure below (Fig. 3) (where we have indicated a possible path) so that each segment links a pair of consecutive numbers, to form, from the beginning to the end the number 1234567?

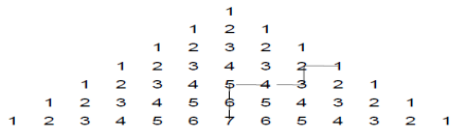

Figure 3. Possible Path

In this session, students were given the problem and asked to describe their approaches to resolving the problem using protocols including: steps in the resolution, explanations of the difficulties they might face, and strategies they would use. Afterwards, each problem was followed by a questionnaire based on the measure of fuzzy rating scales and focused on heuristics related to backwards thinking and the difficulties that are generated during the process of solving problems and emotions and cognitive processes (Gómez-Chacón, in press).

The questionnaire is based on the measure of fuzzy rating scales that used a scale of free fuzzy numbers, in which the respondent represents the same fuzzy number that most closely matches their assessment of an interval (Fig. 4). The questionnaire has two parts, one referring to the cognitive dimension and the other, the affective dimension. The cognitive dimension refers to the characterization of the personal meanings of the subjects on the cognitive dimension of heuristic backtracking, or backward reasoning, and the cognitive appraisal processes of the interaction with

emotion. The studied emotions were: confusion, uncertainty, hesitation, surprise, frustration, bewilderment, boredom, and confidence. And the cognitive appraisal dimensions to differentiate emotional experiences were pleasantness, attentional activity, control (self-other responsibility/control, situational control), certainty, goal-path obstacle, anticipated effort and mental flexibility.

Data analysis

Both qualitative and quantitative methods were used to address the subject of this study. This paper presents the quantitative analysis performed on the undergraduate students' written responses to the questionnaire. The first step of the analysis was the defuzzification of data. This refers to converting the trapezoidal numbers provided at the student's responses into usual numbers that can be handled by the CRT methodology. For the purposes of our study, the average centre (also known as the centroid) defuzzification method was used. With such defuzzicated data, different regression tree analysis were performed with SPSS to uncover the prescriptive nature of the variables. Two of these regression trees, together with their associated rule models, are reported next.

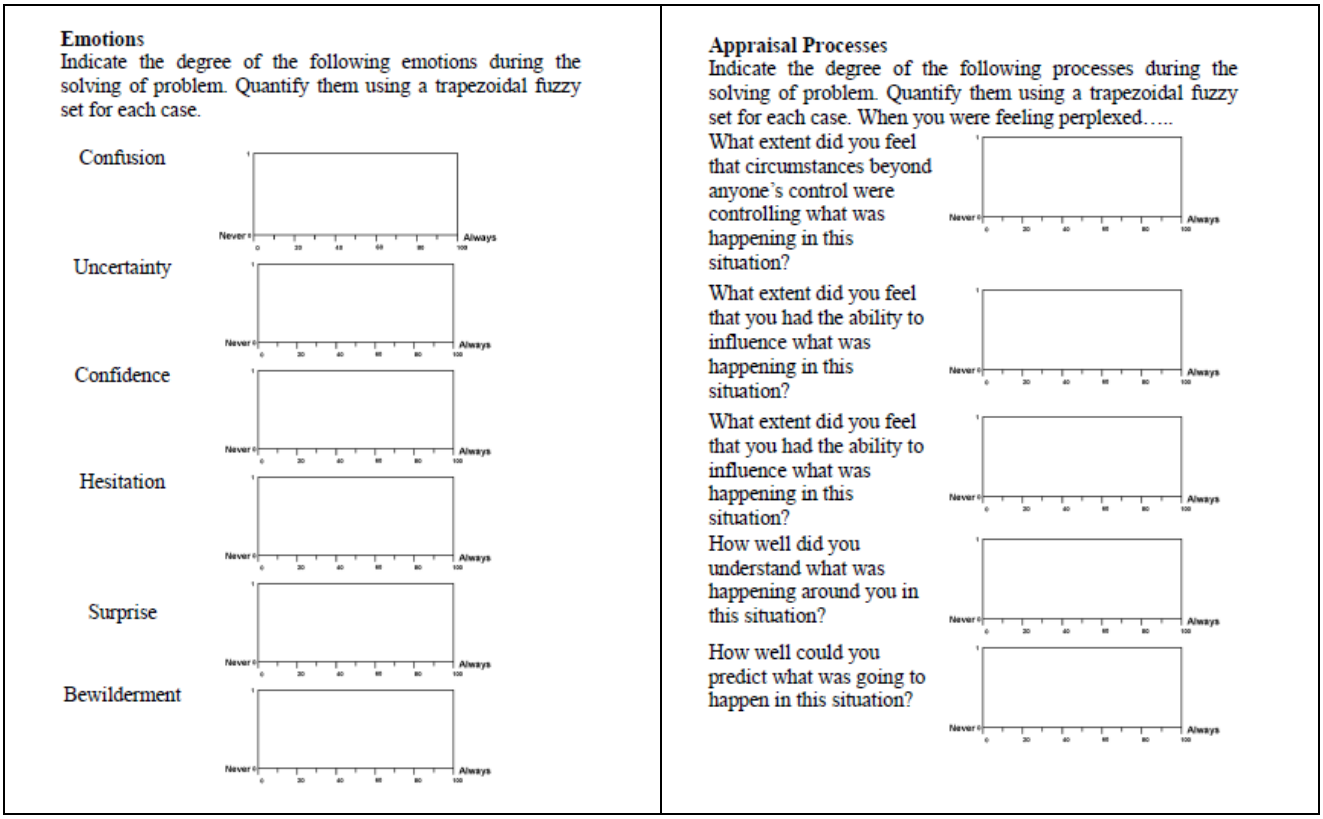

Fig.. 4 Examples of some items from the questionnaire: Perplexity and backwards reasoning processes (Gómez-Chacón, in press)

Results

Research question 1. For the interpretation of the classification tree, we should go looking at the nodes and branching them until the final leaves. First, we look at the root node 0 that describes the dependent variable: Perplexity of students to solve the problem (P2). It indicates that the group mean is 49.040. Then, note that the data is split into nodes 1 and 2 depending on the variable

Bewilderment (P19), indicating that this is the main predictor variable. Node 1 indicates that 22% of students who feel Bewilderment <= 21.37 has a mean of 25.04 perplexity (P2). This node 1 is again split up into nodes 3 and 4 depending on variable P8, Ability to influence (i.e. Control). We note in node 4 that the students who had Control > 67.87 experience perplexity with an average intensity of 40.28, while students at node 3 have a lower ability to influence and experience a mean perplexity intensity of 16.33. These two nodes 3 and 4 are leaves that allow us to infer rules 1 and 2 below. Particularly, each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the conditions along the path to form the antecedent part, and taking the leaf's mean as the rule prediction or consequent. Similarly, in order to define the rest of the rules, node 2 and the following ones are studied. The profile of students who experience perplexity is defined by nodes 3, 4, 5, 9, 10, 11 and 12 through the following variables: Ability to influence (P8), Bewilderment (P19), Confusion (P13), Boredom (P21) and the ability to solve simpler problems and also goal-path obstacles (P11). The inferred rules are the following:

Rule 1 (node 3): IF ((Bewilderment (P19) <= 21.37)) AND (Ability to influence (P8) <= 67.87) THEN the prediction of perplexity (P2) is = 16.33, with a support of 14% (i.e. 14% of the participants verify the premise of this rule).

Rule 2 (node 4): IF ((Bewilderment (P19) <= 21.37)) AND (Ability to influence (P8)> 67.87) THEN the prediction of perplexity (P2) is = 40.28, with support 8%.

Rule 3: (node 5): IF ((21.37 < (Bewilderment (P19) <=64.6)) THEN the prediction of perplexity (P2) is = 48.84, with support 48%.

Rule 4 (node 9): IF (Bewilderment (P19)> 64.6) AND (Confusion (P13) <= 64.87) AND (Boredom (P21) <= 12.62) THEN the prediction of perplexity (P2) is = 67.15, with support 5%.

Rule 5 (node 10): IF (Bewilderment (P19)> 64.6) AND (Confusion (P13) <= 64.87) AND (Boredom (P21)> 12.62) THEN the prediction of perplexity (P2) is = 55.65 with support 5%.

Rule 6 (node 11): IF (Bewilderment (P19)> 64.6) AND (Confusion (P13)> 64.87) AND (the ability to solve simpler problems and goal-path obstacles (P11) <= 80.75) THEN the prediction of perplexity (P2) is = 66.99, with support 15%.

Rule 7 (node 12): IF (Bewilderment (P19)> 64.6) AND (Confusion (P13)> 64.87) AND (the ability to solve simpler problems and goal-path obstacles (P11)> 80.75) THEN the prediction of perplexity (P2) is = 78, with support 5%.

In summary, the perplexity is closely linked with the emotions of bewilderment and confusion. The bewilderment could generate a fork towards a positive path depending on the ability to influence on the problem and the ability to influence the process of resolving. The perplexity state may stem entail only high novelty, reflecting a state of uncertainty but may entail a searching of understanding when perplexity may share with appraisal dimensions linked to the ability to influence (self-control dimension) and the perception of overcome obstacles and the ability to solve simpler problems.

Research question 2. Pleasantness. Pleasantness is considered as an important dimension. It is a function of two appraisals—appraisals of what one wants in relation to what one has, and these are intrinsically pleasant or unpleasant. The mean of the group with respect to pleasure (P5)

experienced during the state of perplexity is 40.92. From the classification' tree (Fig.6) can infer the following rules: **Rule 1 (node 2):** IF Confidence (P15) >55.75 THEN the prediction of pleasure (P5) = 48.07, with support 58%. **Rule 2 (node 3):** IF Confidence (P15) <= 55.75 and Understanding (P9) <= 38.37 THEN the prediction of pleasure = 16.63, with support 13%. **Rule 3 (node 4):** If Confidence (P15) <= 55.75 and Understanding (P9) > 38.37 THEN the prediction of pleasure (P5) = 37.5, with support 29. In summary, a state of perplexity could not only be a mental perturbation or anxiety, but a pleasure experience given sufficient levels of confidence and understanding.

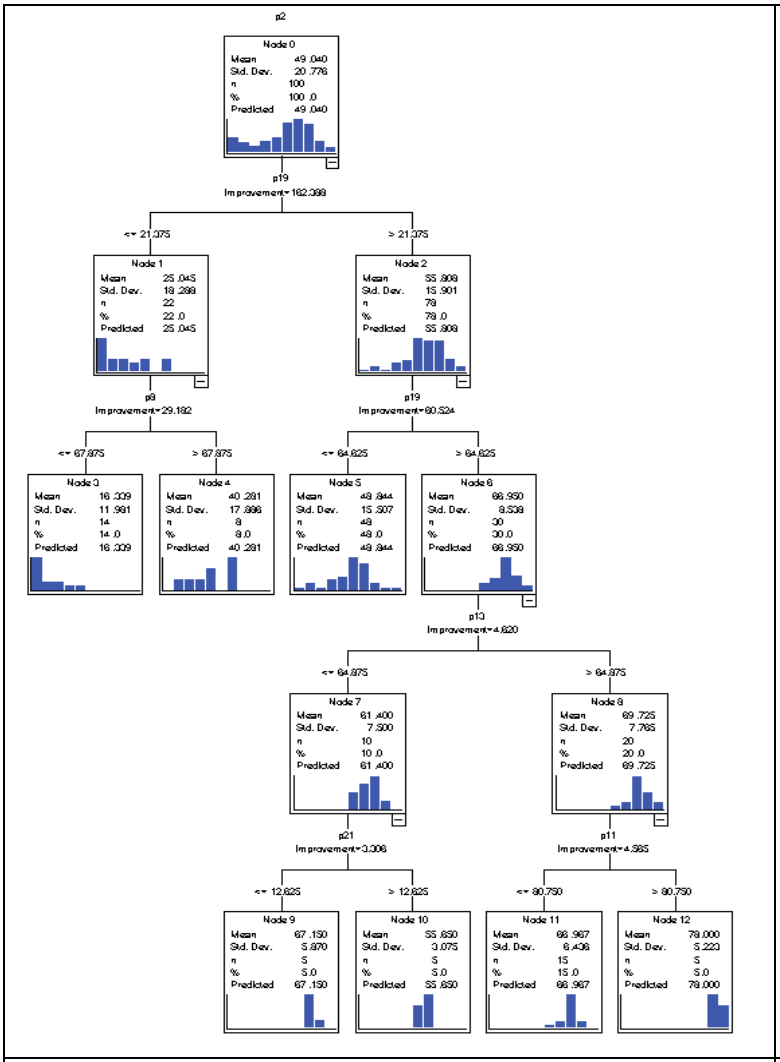

Fig. 5. Regression tree for variable 'Intensity of perplexity' (P2)

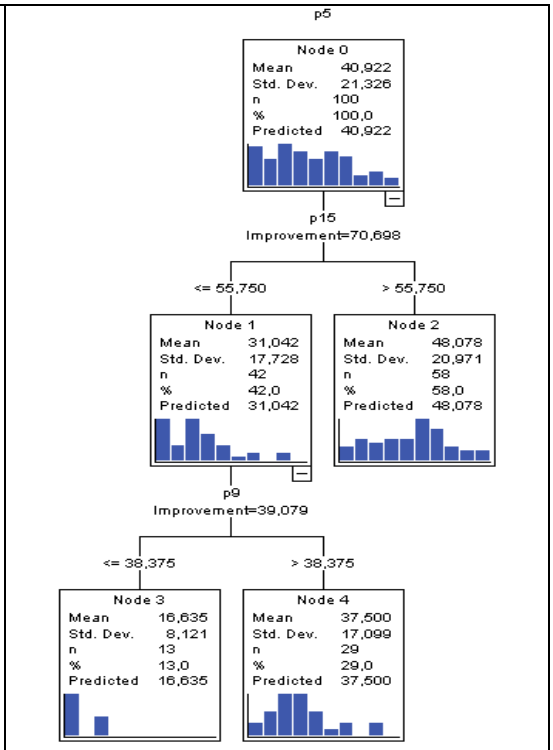

Fig. 6. Regression tree for variable Pleasure (P5).

Conclusions and discussion

The discussion and conclusions are structured around the objectives of the research, and methodological effectiveness in the use of regression trees for establishing rules.

Relative to the interplay between cognition and affect in the perplexity state the perplexity is closely linked with the emotions of bewilderment and confusion. The degree to which students associate state of perplexity with an emotion of pleasure is linked to the levels of confidence and the

understanding of the problem. Likewise the perplexity state shares cognitive appraisal dimensions linked to the ability to influence (self-control dimension) and the perception of overcome obstacles and the ability to solve simpler problems. The relationship shown between cognitive appraisal dimensions and the emotions that make up the state of perplexity highlights conditions about learners who have the ability to appropriately manage their perplexity. This study shows the central role of impasse in mathematics, perplexity it is not a negative event to avoided by intellect, it is responsible for the activation of thinking (Lakatos 1976, Goldin 2000; Gómez-Chacón, 2000). Regarding the methodological adequacy of the present study, the use of a non-parametric, assumptions-free data mining model as regression trees provides a solid basis for the kind of exploratory analysis aimed at this work. Particularly, this model allows for robust variable selection, as significant variables are identified through a greedy process in which the effectiveness of all the available variables in reducing the variability of the response variable is checked, and that obtaining the greatest reduction (or improvement) is selected. This assures that the selected variables that conform the premises of the obtained rules are relatively good explicative factors of the studied response variables.

Acknowledgments: The paper' elaboration has been supported by special action grant TIN2015-66471-P from the Government of Spain and the research grant Visiting Scholar Fellowship, University of California in Berkeley, Scholarship "Becas Complutense del Amo" – Spain.

References

Breiman, L., Friedman, J. H., Olshen, R. & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, California.

Gil, M. A., Lubiano, M. A., de la Rosa de Sáa, S., & Sinova, B. (2015). Analyzing data from a fuzzy rating scale-based questionnaire. A case study. *Psicothema*, 27(2), 182–191.

Goldin, G. A. (2000). Affective pathways and representations in mathematical problem solving. *Mathematical Thinking and Learning*, 17(2), 209–219.

Gómez-Chacón, I. M. (2000). Affective influences in the knowledge of mathematics. *Educational* Studies in Mathematics, 43, 149–168.

Gómez-Chacón, I. M. (in process). Emotions and Heuristics: The state of perplexity in mathematics, *ZDM-The International Journal on Mathematics Education*

Lakatos, I. (1976). *Proofs and Refutations*. Cambridge, UK: Cambridge University Press.

Silvia, P. J. (2010). Confusion and interest: The role of knowledge emotions in aesthetic experience. *Psychology of Aesthetics Creativity and the Arts*, 4, 75–80.

Smith, C. A., & Ellsworth, P. (1985). Patterns of cognitive appraisal in emotions. *Journal of Personal and Social Psychology*, 84(4), 813–838.

Strobl, C., Malley, J., and Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological methods*, 14(4), 323–348.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences* 8, 199-249.