Journal of Information Technology Education: Innovations in Practice

An Official Publication
of the Informing Science Institute
InformingScience.org

JITEiip.org

Volume 18, 2019

CONSTRUCTED RESPONSE OR MULTIPLE-CHOICE QUESTIONS FOR ASSESSING DECLARATIVE PROGRAMMING KNOWLEDGE? THAT IS THE QUESTION!

Yolanda Belo | Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal | yobelo97@gmail.com

Sérgio Moro | Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal | scmoro@gmail.com

António Martins | Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal | sigforma@gmail.com

Pedro Ramos | Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal, and Instituto Universitário de Lisboa (ISCTE-IUL), IT-IUL, Lisboa, Portugal | pedro.ramos@iscte-iul.pt

Joana Martinho Costa* | Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal | joana.martinho.costa@iscte-iul.pt

Joaquim Esmerado | Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal | joaquim.esmerado@iscte-iul.pt

* Corresponding author

ABSTRACT

Aim/Purpose: This paper presents a data mining approach for analyzing responses to advanced declarative programming questions. The goal of this research is to find a model that can explain the results obtained by students when they perform exams with Constructed Response questions and with equivalent Multiple-Choice Questions.

Background: The assessment of acquired knowledge is a fundamental role in the teaching-learning process. It helps to identify the factors that can contribute to the teacher in the developing of pedagogical methods and evaluation tools and it also contributes to the self-regulation process of learning. However, better format of questions to assess declarative programming knowledge is still a

Accepting Editor Dennis Kira │ Received: October 30, 2019 │ Revised: December 9, December 18, 2019 │ Accepted: December 21, 2019.

Cite as: Belo et al. (2019). Constructed response or multiple-choice questions for assessing declarative programming knowledge? That is the question! *Journal of Information Technology Education: Innovations in Practice, 18,* 153-170. https://doi.org/10.28945/4479

subject of ongoing debate. While some research advocates the use of constructed responses, others emphasize the potential of multiple-choice questions.

Methodology A sensitivity analysis was applied to extract useful knowledge from the relevance of the characteristics (i.e., the input variables) used for the data mining process to compute the score.

Contribution Such knowledge helps the teachers to decide which format they must consider with respect to the objectives and expected students results.

Findings The results shown a set of factors that influence the discrepancy between answers in both formats.

Recommendations for Practitioners Teachers can make an informed decision about whether to choose multiple-choice questions or constructed-response taking into account the results of this study.

Recommendations for Researchers In this study a block of exams with CR questions is verified to complement the area of learning, returning greater performance in the evaluation of students and improving the teaching-learning process.

Impact on Society The results of this research confirm the findings of several other researchers that the use of ICT and the application of MCQ is an added value in the evaluation process. In most cases the student is more likely to succeed with MCQ, however if the teacher prefers to evaluate with CR other research approaches are needed.

Future Research Future research must include other question formats.

Keywords constructed response, multiple-choice questions, educational data mining, support vector machine, neural networks

INTRODUCTION

Assessment is one of the critical components of the educational process. When it is used in an appropriate way, it can be a decisive factor for achieving the objectives of the subject (Camilo & Silva, 2008). Thus, there are several test models, from the most traditional paper-based ones up to electronic format, composed by questions requiring Constructed Response (CR) where they are directly asked (Clark, 2004), and for Multiple-Choice Questions (MCQ) with the presence of several alternatives where only one is correct, in true-false format, open space and other formats to express this type of test (Pinto, 2001). Therefore, the present study intends to verify whether the students answers to MCQ or CR questions, obtain similar results.

For example, Kuechler and Simkin (2003) consider that, since most teachers have a greater preference for CR over MCQ, the fact is that students with a high level of performance in the subject must assess the questions themselves, a task which is more protracted than MCQ which requires more subjectivity (Zeidner, 1987, cited by Kuechler & Simkin, 2003). On the other hand, it is necessary to question how converging is the commitment of the student in MCQ in relation to the commitment of CR. For these reasons, for many years researchers have tried to respond to these types of questions, in a way that takes advantage of the teaching-learning process in its entirety, both for students as well as teachers.

This study aims to explain the possible discrepancy between the two types of questions. The questions are paired, that is, a topic of the subject is tested by the two assessment methods in the same test. A sample of tests from the academic year 2016/2017 on declarative programming courses is used.

Data Mining (DM) enables the extraction of meaningful patterns from data which can be translated into actionable knowledge. As such, it has been applied to educational data to provide novel insights grounded on large sets of data, leading to a new trend named Educational Data Mining and/or Learning Analytics (Baker, 2010; Baradwaj & Pal, 2011; Slater, Joksimović, Kovanovic, Baker, & Gasevic, 2017). In this study, DM was adopted to create an explanatory model (with a dataset composed of data collected from a sample of 300 Excel Advanced assessment tests in which 50% are open and 50% are MCQ) illustrating relevant attributes (i.e., input variables which reflect the characteristics of questions and students) to the study, the implication of each one in the results, and to confront differences in both assessment methods. The contributions of this study may help the teacher in choosing the evaluation format, and if the model chosen is adapted to the intended teaching-learning process, whether in CR or in MCQ formats.

BACKGROUND

TAXONOMIES FOR CONSTRUCTION OF QUESTIONS

It is important to consider the clear and structured definition of educational objectives, since the acquisition of knowledge and skills appropriate to a professional profile to be acquired should be directed from a teaching process with adequate choices of strategies, delimitation of specific contents, assessment tools, and consequently lead to effective and lasting learning. This requires a typology of processes and objectives of learning, to help the identification, declaration and control of educational objectives linked to a set of processes from the acquisition of knowledge, skills and attitudes, to the planning of teaching and learning (Ferraz & Belhot, 2010).

There is a set of suitable taxonomies for creating questions about declarative programming knowledge, like the SOLO Taxonomy (Structure of the Observed Learning Outcome) that assesses the responses from two dimensions: the level of abstraction and the increased complexity of performance across tasks (Biggs, 1996); the Matrix Taxonomy that separates the ability to produce and interpret programming code (Füller et al., 2007); and the Bloom's Taxonomy, a taxonomy of educational objectives that categorizes the questions according to the level of cognitive domain complexity required for their resolution (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1976). Subsequently, the Bloom's Taxonomy was revised by Krathwohl (2002). In this version the changes essentially consist of the organization of knowledge in a matrix rather than a one-dimensional presentation to allow the distinction between the type of knowledge and the cognitive process, and the replacement of the names of each category with verbs that approach the names of learning objectives.

Bloom's Taxonomy is a taxonomy of educational objectives that allows a hierarchically structuration of the questions by levels, from the simplest and most concrete to the most complex and abstract (Krathwohl, 2002; Lister, 2003). Its application assumes that students reach a new level only when they complete the previous level. This taxonomy allows the classification of learning objectives so that the teacher can accurately verify the level achieved by each student. Course planning is also used as it allows curricula to be organized according to the objectives to be achieved (Bloom et al., 1976).

Bloom et al. (1976) report that it is expected that more success is achieved at the early levels of the taxonomy and there will be a decrease moving up the hierarchy. The authors also presented a study in which they found that subjects with lower school outcomes are more likely to achieve lower outcomes at higher taxonomy levels and higher outcomes at lower levels than the reverse situation. Simkin and Kuechler (2005) found there is difficulty in constructing MCQ that reach a high level of learning in comparison to CR when referring to the application of Bloom's Taxonomy. In addition, the research concludes that, the results tend to be positive in MCQ if they are developed around the understanding level, and CR respectively for application to higher ones. However, Buchweitz (1975) compared the results of MCQ with those of CR, concluding that there is no significant difference between evaluating by the first type or the other, even for all educational levels of Bloom's Taxonomy. In addition, the first four levels of the taxonomy (knowledge, understanding, application and analysis) can be applied in MCQ format exams, while the last two (synthesis and evaluation) would be better evaluated in essay questions format, not discarding the possibility of being also indicated for the previous levels (Costa & Miranda, 2017; Gronlund, 1988).

Generally, CR have been the preference over MCQ by several educators, based on the belief that the first method measures a greater number of skills on students' comprehension and ability, while MCQ reflects less cognitive aspects regarding the application of knowledge and the art of producing a response (Chan & Kennedy, 2002). Although there is a frequent trend on students to preferring MCQ over CR because of their apparent ease (Pinto, 2001). For Lukhele, Thissen, and Wainer (1994), there is no clear standard of possible answers in CR format, and there is great difficulty in interpretation and subjective compilation by the teacher, CR carries limitations to ensure a uniform and quality evaluation, especially if there are several teachers. In this way, according to Čandrlić, Katić, and Dlab (2014) with online tests, there has been a transition from evaluations on paper - based and using CR, to electronic MCQ tests as their evaluation tools. In addition, the use of MCQ in evaluations are easy to apply and analyze because they do not require elaborate student responses as happens in CR, MCQ offers quick response (depending on the difficulty of the question), and consequently can be objectively registered and classified by the teachers (Pinto, 2001).

MCQ and CR have been studied in several areas such as Mathematics (e.g., Katz, Bennet, & Berger, 2000; Stankous, 2016) and Economics (e.g. Hickson, Reed, & Sander, 2010; Kennedy & Walstad, 1997), highlighting potentials and disadvantages of both types of questions. However, none of these studies was peremptory in discouraging the use of MCQ. In Computer Science a set of studies argues that MCQ should be used in this area. Roberts (2006) used the MCQ in his study to evaluate undergraduate students for enhancing the process of learning. He concludes from students' feedback that the use of MCQ for evaluations seems to be an effective way of learning. However, he highlights that the conclusions of his work must be researched further with more data. Kuechler and Simkin (2003) studied the correlation between MCQ and CR questions in a programming course and found small differences between both, only affected by gender, dummy values and coding. From their results and the characteristics of MCQ, the authors conclude that this type of questions should be preferential when compared to CR questions.

Excel Learning and Mining Educational Data

Spreadsheets are the most used declarative programming language application (Burnett et al., 2001). They are commonly used in accounting, health, marketing; and in areas requiring a little more programming, such as engineering, where a set of design activities, documentation, debugging, testing, maintenance, storage and computation exist (Maresca, 2016). Thus, in an educational area that involves a combination of practical knowledge and abstraction, using the computer and Excel, Silva (2009) states that there is a contribution to the establishment of an educational process that allows both the student to understand about the importance of knowledge as a new process of evaluation that allows the replacement of calculator, paper and pencil. Where the student is faced with situations and problems, they will learn to develop strategies that acquire the spirit to research, experiment, data organize, systematize the results, validate the solution, as well as the expansion of new knowledge.

Within the studies by Almeida (2017) and Cortez and Silva (2008), where the aim was the identification of the factors that influence the success of a student, in exams of Advanced Excel and Introduction to Excel, and the prediction of the student's results with the identification of the factors that influence educational success/failure in Mathematics and Portuguese classes, by applying DM techniques: MLPE (Multilayer Perceptrons), SVM (Support Vector Machines), DT (Decision Trees), NB (Naïve Baies) and other techniques respectively. Educational Data Mining has been considered a research area that is concerned with the search for methods that explore educational data, in which exist an objective of perceiving students and their academic performances, as well as to explore better ways of learning. Therefore, it was possible to conclude from these studies that the examinations that had a very long MCQ enunciation are one of the main causes that can influence negatively the results obtained by these questions, either by the student's interpretation or even misunderstanding of the objective of the question, or the degree of difficulty and the topic of the subject. It is also possible to predict student outcomes, especially when associated with social and educational factors.

We wanted to verify what were the most predominant factors pointed out by the literature in the CR and MCQ students' answers about declarative programming knowledge. A sensitivity analysis was applied to extract useful knowledge from the relevance of the attributes against the score which will be described in the sections below.

MATERIALS AND METHODS

DATA COLLECTION AND PREPARATION

The empirical experiments were based in hand-written Excel exams performed at ISCTE-IUL, in the academic year 2016/2017. They were composed on Excel´s objective formulas (see Appendix A) and the basic structure of the exam consisted of two blocks: the first with 10 CR questions and the last by 10 paired MCQ. Regarding CR, the student gets one (1) point for each correct answer, zero (0) for incorrect and a grade on a scale of zero to one depending on what was expected. For MCQ, it was possible to identify three possible scoring cases, one for correct, 0 for unanswered and 0.25 discount for each wrong answer. The dataset compiled includes a total of 2787 records corresponding to the students' responses in each question from the exam.

Appendix B shows all the attributes included in this study, including one feature that keeps the difference between CR and MCQ scores. Note that, we chose to use difference in real values (instead of absolute) for the importance that the variable can bring with more details for the research. The number of variants indicates that the classes would not repeat the same exam, however the structure still the same. The degree of difficulty had as a criterion the composition of operations/formulas (see Almeida, 2017). The question topic was considered because it would be interesting to verify the student's performance with the content type of the question (see Hudson, 2012). We also classified each question with their level of Bloom's Taxonomy, through the required skills and behaviors according to the objectives, like Scouller (1998). The student gender was also considered to evaluate whether the scores are equal or not, as happens with Hudson (2012).

DATA MINING MODELS AND KNOWLEDGE EXTRACTION

Data Mining (DM) encompasses the process of data visualization, with the objective of automatically inferring models and rules that have an implicit knowledge of the data studied (Quintela, 2005). In this study, CRISP-DM (a Data Mining methodology) was chosen because it is one of the most used and widely accepted methodologies, as well as having extensive literature available on the methodology (e.g., Moro, Esmerado, Ramos, & Alturas, 2019). By applying DM, it is possible to train a model that reflects the different features (i.e., variables) that characterize the problem. Since the goal is to model the grade achieved by a student for a given question, it becomes a supervised learning problem. Specifically, the fact that the grade is a numeric value turns the problem into a regression one (from the DM perspective). Thus, DM enables the modeling of a problem by developing a computational model that attempts to predict the outcome feature given a set of values for a list of the input features (Moro et al., 2019).

There are several DM techniques that can be applied to a regression problem. In this case, we adopted Decision Trees (DT), Random Forest (RF), the Neural Networks (NN) on their variances

MLP (Multilayer Perceptron) and MLPE, K-NN (K-Nearest Neighbors) and finally the Support Vector Machines (SVM), considering the scientific work of Fernandes, Moro, Costa, and Aparício (2019), with mechanisms to measure the estimation of the error, MAE (Mean Absolute Error) and NMAE (Normalized MAE) (more information at Silva, Moro, Rita, and Cortez, 2018). DT consists of a structure that connects a set of nodes through branches resulting from a recursive partition of the data, from the root node to the leaves, each branch representing a conjunction of conditions, as well as the leaves (pure nodes) correspond to classes, internal nodes to attributes, and branches to attribute values (Quintela, 2005). The RF model is based on building a series of DT and use them in combination, but it cannot be directly interpretable as is possible for an individual DT. Although with the RF model it is still possible to provide explanatory knowledge in terms of its input variable relevance (Cortez & Silva, 2008). NN attempts to mimic the complexity of the human brain through a model constituted by nodes (or neurons) and connections between them (or synapses). The complexity of the model comes from the number of nodes and their connections (Quintela, 2005). During the learning process, the NN, through a learning or training algorithm, adjusts the connection weights until a satisfactory result is achieved. On the other hand, the SVM are techniques considered "black box" (as NN also), that is, the extraction of knowledge is encoded in equations with difficulty on interpreting. (Quintela, 2005).

The main activities performed in the Modeling phase and Evaluation are graphically illustrated in figure 1. From division of tests and training, application of prediction metrics, to the extraction of knowledge which helps decision making and business processes. The data-based sensitivity analysis (DSA) enables the capture of input features' influence on the output feature by assessing how much the output changes as a result of changing simultaneously a set of randomly selected values for the input features (Cortez & Embrechts, 2013).

The experiments were performed using the open source R statistical tool (with the rminer package), installed in the R environment (RStudio v 1.1.423), like Almeida (2017) and Moro et al. (2019) to build models for pattern trainings, explaining the influence of each attributes to initial objective and finally to extract knowledge. Note that, to ensure robustness the dataset was divided into K=10 parts of 10 runs (k-fold validation scheme). Next, the same error metrics were applied to all models, so no error was inserted that would impair the comparison of the results. The best results were obtained by SVM with MAE by 31%, while NMAE, by 18%, was the model with better accuracy. Figure 2 plots the regression error characteristic (REC) curves for the six tested models. REC is generalization for regression problems of the receiver operating characteristic (ROC) curves used to assess a classifier's performance. "The ROC curve characterizes the performance of a binary classification model across all possible trade-offs between the false negative and false positive classification rates" (Bi & Bennett, 2003; p. 43). Likewise, the REC curve enables assessment of the trade-off between error tolerance versus the accuracy of the models. Thus, the higher the distance from the curve to the imaginary diagonal line, the better the model.

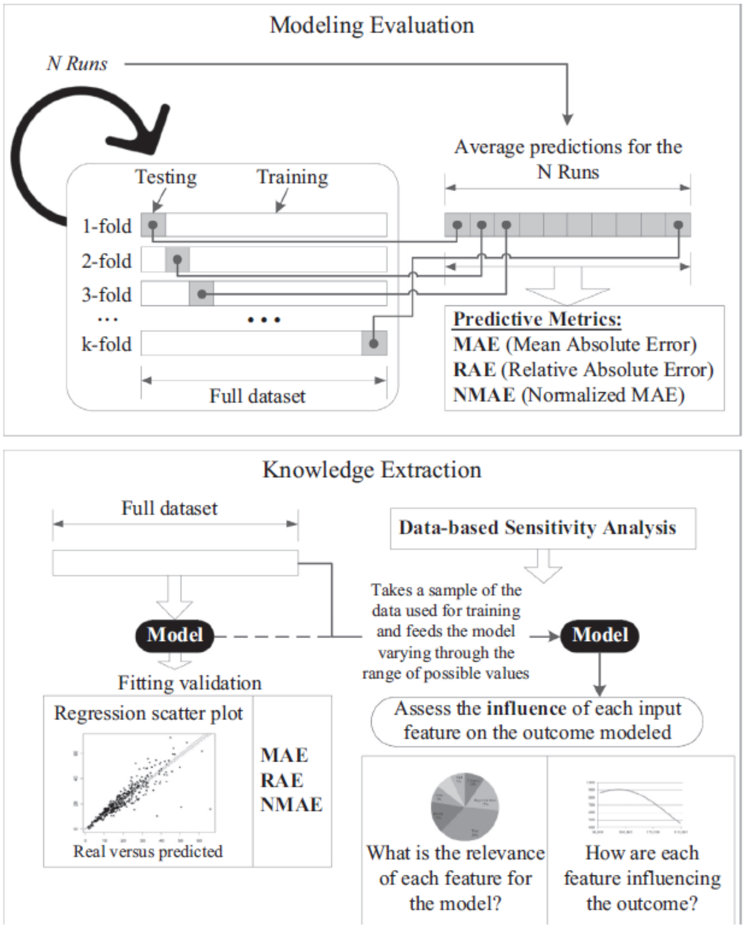
Modeling Evaluation
N Runs
Testing
Training
1-fold
2-fold
3-fold
…
k-fold
• • •
Full dataset
Average predictions for the N Runs
Predictive Metrics:
MAE (Mean Absolute Error)
RAE (Relative Absolute Error)
NMAE (Normalized MAE)
Knowledge Extraction
Full dataset
Model
Fitting validation
Regression scatter plot
Real versus predicted
MAE
RAE
NMAE
Data-based Sensitivity Analysis
Takes a sample of the data used for training and feeds the model varying through the range of possible values
Model
Assess the influence of each input feature on the outcome modeled
What is the relevance of each feature for the model?
How are each feature influencing the outcome?

Figure 1. Modeling and Evaluation Activities

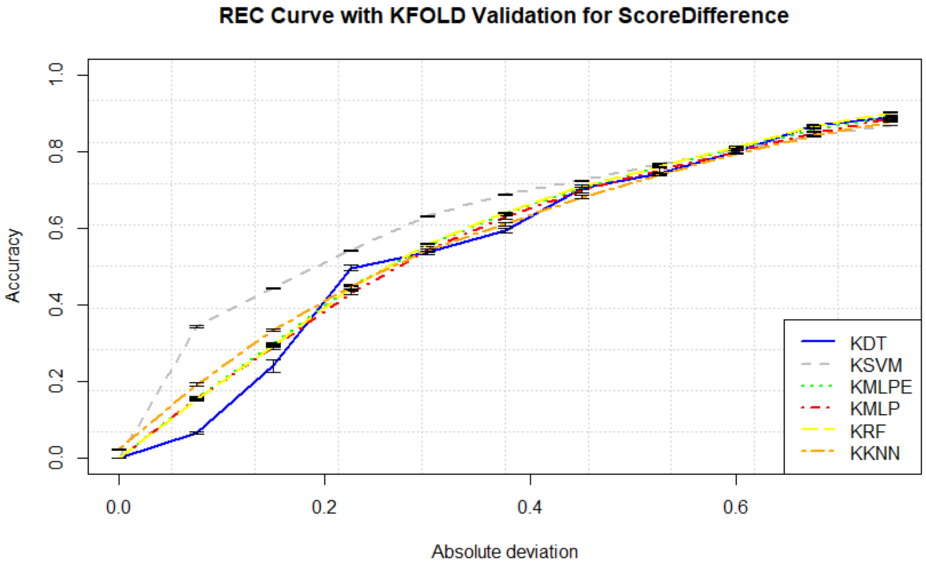

Figure 2. REC Curve

RESULTS AND DISCUSSION

When dealing with black box models, it is often challenging to extract knowledge. Consequently, DSA (Data-based Sensitivity Analysis) have emerged to deal with this problem (Saltelli et al., 2000 cited by Silva et al., 2018), identifying the relevant features to the model and their influence in descending order of importance, as shown in Figure 3. Because the attributes were relatively close to one another, they were all considered for knowledge extraction, and to explain the assumption that students have a better chance of succeeding in MCQ. Although some of them revealed values approaching the zero point (proposing no difference in the exams format or benefiting CR).

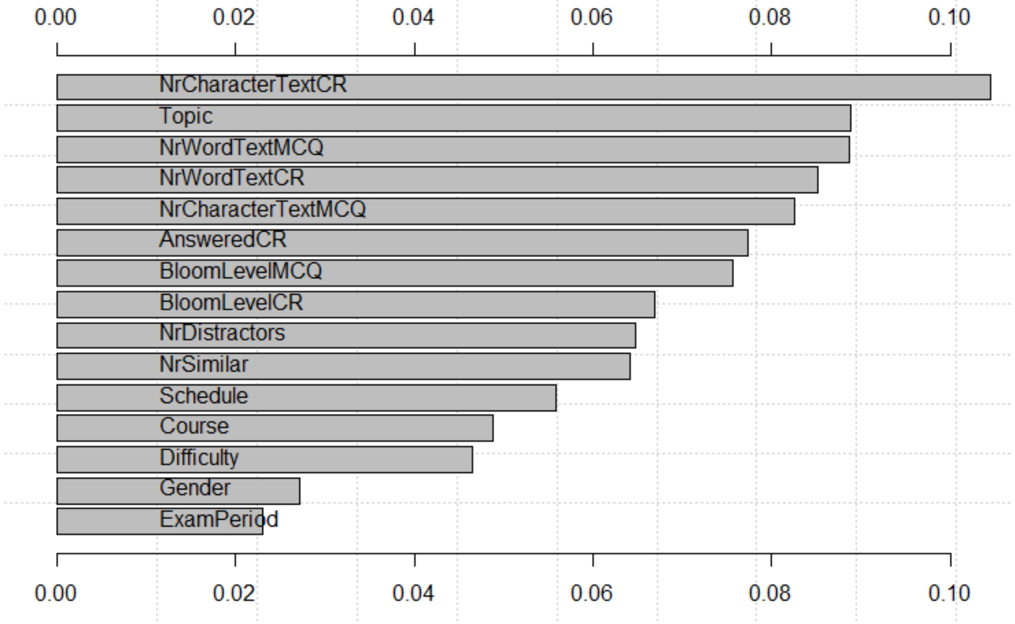

Figure 3. Features Relevance

Next, the input features used to train the model are scrutinized through variable effect characteristic (VEC) graphics, which are drawn upon the results of the DSA, as detailed by Cortez and Embrechts (2013). The VEC plots how a given input feature affects the output, in this case, the difference in the students' score between both types of questions.

The most relevant feature indicates how long the CR question text is. Thus, Figure 4 indicates the higher the number of characters in CR questions, the higher the probability of the student to fail on this format and to succeed on MCQ (like Santos et al., 2011 cited by Almeida, 2017). The Topic feature, with 9% importance, discriminates the influence they have on the results in the two formats. So, for Statistics this affects the variable target indicating success in CR. Unlike for example the Logical whose influence assumes a value further from the zero point and therefore, students are more likely to succeed in MCQ. According to Almeida (2017), the topic of the question is a very important factor, since the teacher can obtain a sense of which topics of the subject the students have more difficulties, and therefore where they can be better applied, whether in CR or MCQ exams.

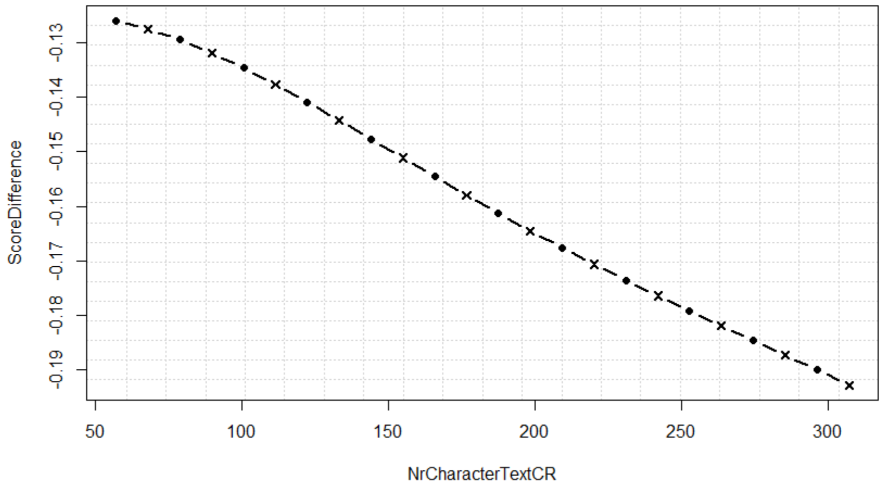

Figure 4. NrCharacterTextCR and ScoreDifference

According to the results shown on Figure 5, the larger the length of the MCQ text, the higher the probability of being misinterpreted by the student (Almeida 2017), and as expected, the higher the difference in both formats providing better achievements in CR questions. Nevertheless, Figure 6 demonstrates an opposite scenario in which, the higher the number of words in the CR text, the higher the probability that both formats are equal or, that students will be more successful in CR as the number of words in the text grows, in contrast, the higher the number of characters in the CR text, the more likely the student score in MCQ. Considering an unexpected situation since a word is a combination of characters, so the interpretation of both should be similar. Unless we consider that in CR questions where the text has a greater amount of excel functions, the students consider it complicated and fail, whereas, in CR with fewer excel functions, they are more likely to succeed. As stated by Dubins et al. (2016) cited by Almeida (2017), the misinterpretation factor of the text can be considered as a factor for the student not answering the question, since it may be associated with an incorrect reading or difficulty in interpreting a poorly worded question.

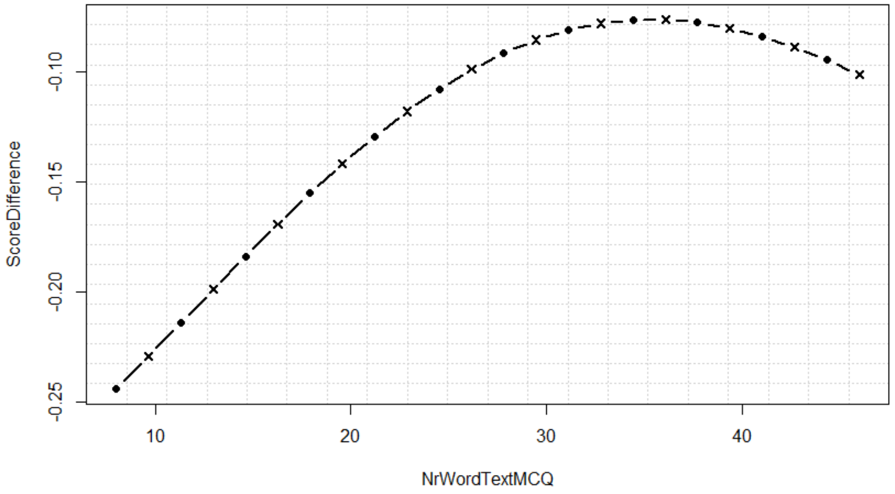

Figure 5. NrWordTextMCQ and ScoreDifference

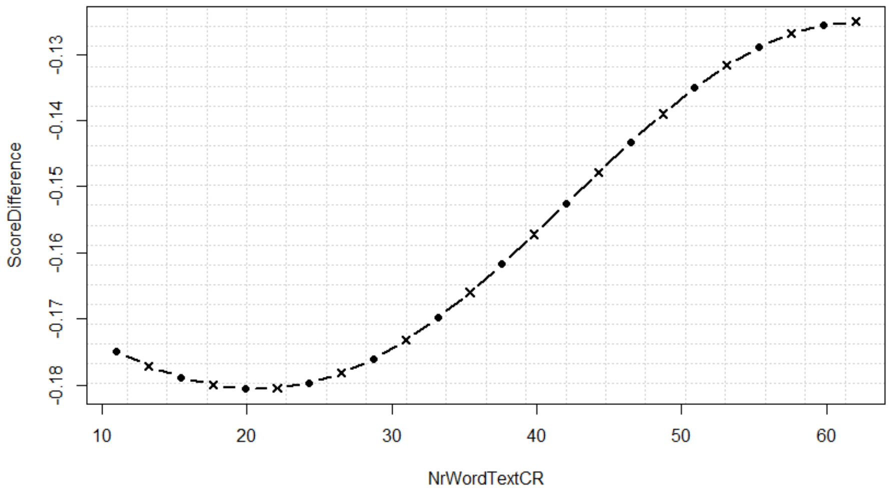

Figure 6. NrWordTextCR and ScoreDifference

The seventh most relevant feature indicates the student's level of learning regarding educational objectives of the Bloom's Taxonomy. In MCQ, the student is more likely to be successful in Analysis, Applying, and Remembering levels (despite that they are presented as values very close to zero) and the Understanding level, as shown on Figure 7. Unlike for CR, the Evaluating level suggests either there is no difference in both formats or higher results in CR questions (see Figure 8). According to Gronlund (1988), the last two Bloom's objectives would be better evaluated in CR questions, and the first four could be applied in MCQ format. Note that for Remembering it remains the same in both formats, that is, for both MCQ and CR, this level remains at an average score of -0.11, still promoting MCQ as the format in which students return better results. These results are also supported by, Gronlund's (1988), Pinto's (2001), and Buckles and Siegfried's (2006) works when they argue that the first four levels of Bloom's Taxonomy, can be properly evaluated with MCQ.

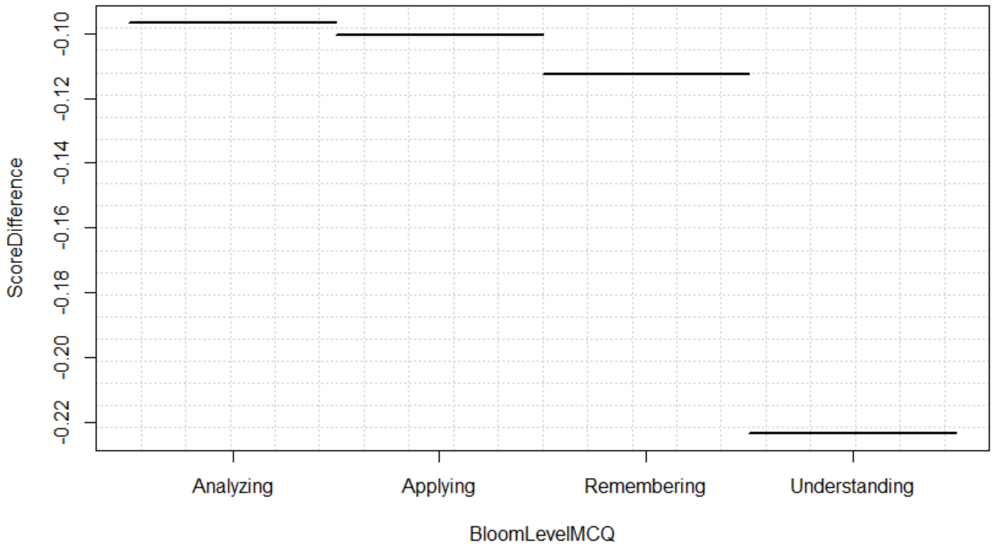

Figure 7. BloomLevelMCQ and ScoreDifference

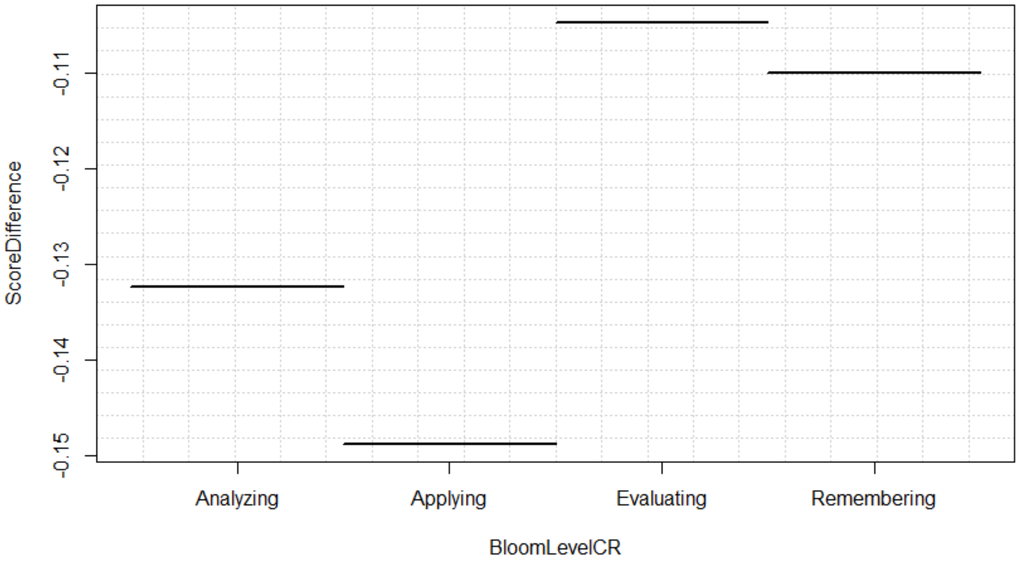

Figure 8. BloomLevelCR and ScoreDifference

The questions with two distractors benefit the MCQ format, opposite for example when questions with three distractors (or two similar) tend to approximate the model to the zero point (Figure 8). According to Dubins et al. (2016) cited by Almeida (2017), the more similar options are to the correct answer, the higher the probability of success by guessing; and the less options close to the correct, the higher the probability that student settle the question with less hesitation in "risking". Afterwards, the schedule in which the student learns, where the exams performed by students in Daytime benefits CR format (although still indicate better results in MCQ) unlike in the Evening period, where the student succeed more in MCQ.

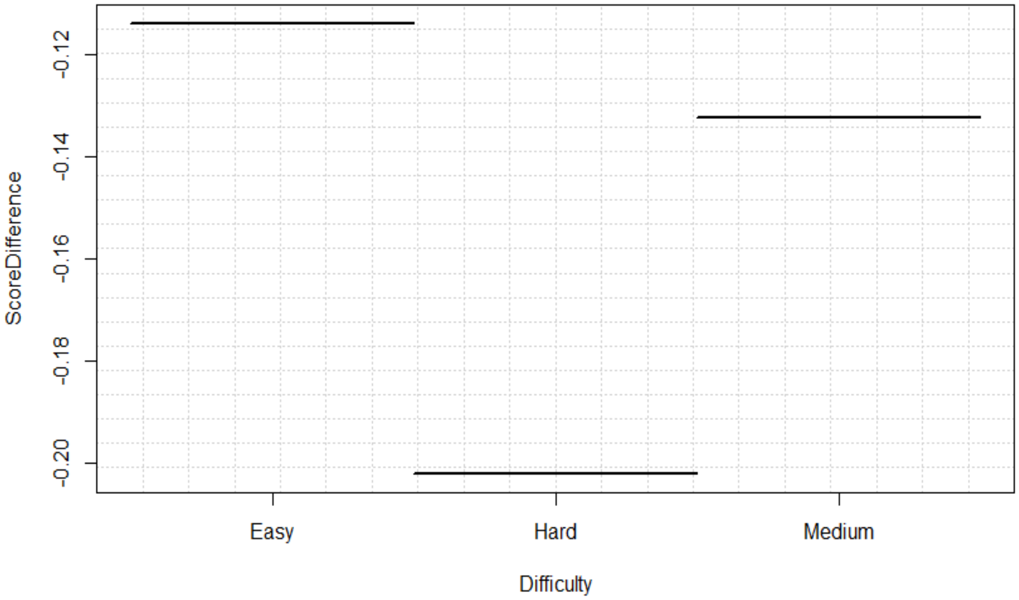

Figure 9. Difficulty and ScoreDifference

The students who attended the Institutional Course (IC) had higher probability of succeed in MCQ, followed by Computer Science and Business Management (CSBM), Telecommunications and Computer Engineering (TCE) and Anthropology (A). On the other hand, this model suggests that students attending Computational Engineering (CE) are more likely to succeed in CR. Considering the difficulty of the questions, the students of the CE course succeed more in MCQ when the difficulty of the questions is Medium. In addition, by Figure 9 it is possible to verify that the hard level is the one in which the student has higher probability to succeed in MCQ, contrary-wise for Easy. Note that, the difference of the results in both formats tends to approach zero, assuming there is no notable difference in the student's results which is similar to the work of Kuechler and Simkin (2003) when they argue that there are small differences between the answers from both types of question. Nevertheless, there are more answers in our results that are correct in MCQ.

CONCLUSION

Much has been discussed recently concerning the teaching and learning process, particularly how educational objectives are defined to guarantee the acquisition of knowledge and competence for a student. The assessment is one of the most used strategies to measure theoretical and practical performance. Moreover, it is important to mention that apart from the objectives included, the format of the test is also part of the assessment method. From all the possible, various exam formats, this paper included only multiple-choice and constructed response formats.

When implementing Data Mining algorithms, the SVM model had better performance, with the calculations using 10-fold, it was possible to verify that both exams are not equal. The students have a better chance of success in exams with a Multiple-Choice Question (MCQ) format. Stankous (2016) argues that, despite the great potential of using MCQ in the mathematics subject, the CR questions assess the students' real knowledge more accurately. In the economic subject, Hickson, Reed and Sander (2010) stated that the difference between the scores of both types is not significant.

In this study, we identified the main factors that influenced this discrepancy between the formats in a computer science subject, specifically on declarative programming knowledge. Firstly, the longer the text of the question, the higher the probability of the student to fail with MCQ questions and to succeed in the equivalent question in the CR format. Haladyna et al. (2012) cited by Almeida (2017) stated that the elaboration of questions requires efforts, and the teachers must elaborate questions as clearly as possible, avoiding long texts. On the other hand, the question topic, and difficulty level were tested, returning results that helps the teacher in identifying which to subjects the student is most likely to respond easily, as well as the complexity of the question. The difficulty of the question cannot only be measured by the number of formulas but also considering distractors or similar answers to the correct one.

Using Bloom's Taxonomy, Simkin and Kuechler (2005) found that, results tend to be more positive in MCQ if these are developed at the Understanding level, and CR are better in the Application level, confirming the results of this research. Thus, the teacher will know which procedures can help in constructing these types of questions, as the requirements of each level differ; and how the student can succeed, resulting in reaching high levels of performance for both teacher and student.

Since several researchers determined that the use of ICT is an added value in the evaluation process with the application of MCQ (e.g. Azevedo, 2017; Scouller, 1998), then it is clear that the results of this research confirm this premise, since in most cases the student is more likely to succeed in MCQ. However if the teacher prefers to evaluate in CR, which was also verified in studies in economics and mathematics subjects (e.g. Stankous, 2016; Hickson, Reed & Sander, 2010), it requires more research on the implementation of exams with both question formats, which is the method that students consider fairer (Hickson, Reed & Sander, 2010).

This study presented research in which the block of exams with CR questions can be verified (since one of the examples of the MCQ block was studied by Almeida, 2017) to complement the area of research returning greater performance in the elaboration of more suitable evaluation to the students and increasing the teaching-learning process. For future work, it is desirable to understand how short- and long-term memory and reasoning are related to the answers in MCQ and to verify if the results will be the same in other scientific areas (e.g. Katz, Bennett & Berger, 2000; Stankous, 2016).

References

Almeida, D. P. (2017). *Fatores de sucesso na avaliação de questões de escolha múltipla: O caso de exames de Excel* [Success factors in assessing multiple choice questions: The case of excel exams]. Masters Dissertation. Lisbon, Portugal: ISCTE – Instituto Universitário de Lisboa. Retrieved from https://repositorio.iscte-iul.pt/handle/10071/15055

Azevedo, J. M. (2017). *Avaliação sumativa em matemática no ensino superior com recurso a questões de escolha-múltipla: Uma abordagem utilizando a metodologia investigação-ação* [Summative assessment in mathematics in higher education using multiple choice questions: An approach using the action-research methodology]. Doctoral Dissertation. Covilhã, Portugal: Universidade da Beira Interior. Retrieved from https://ubibliorum.ubi.pt/handle/10400.6/4493

Baker, R. S. J. D. (2010). Data mining. In *International encyclopedia of education* (3rd ed.) (pp. 112-118). Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.01318-X

Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, *2*(6), 63-69. https://doi.org/10.14569/IJACSA.2011.020609

Bi, J., & Bennett, K. P. (2003). Regression error characteristic curves. *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (pp. 43-50). Retrieved from https://www.aaai.org/Papers/ICML/2003/ICML03-009.pdf

Biggs, J. M. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347-364. https://doi.org/10.1007/BF00138871

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1976). *Taxionomia de objetivos educacionais: Domínio cognitivo*. Porto Alegre: Editora Globo.

Buchweitz, B. (1975). Testes de múltipla escoha e de resposta livre em física geral [Multiple choice and free response tests in general physics]. In *Reunião anual da SBPC (xxvi)* (pp. 3-6). Retrieved from http://publicacoes.fcc.org.br/ojs/index.php/cp/article/viewFile/1784/1767

Buckles, S., & Siegfried, J. J. (2006). Using multiple-choice questions to evaluate in-depth learning of economics. *The Journal of Economic Education, 37*(1), 48-57. https://doi.org/10.3200/jece.37.1.48-57

Burnett, M., Atwood, J., Djang, R. W., Reichwein, J., Gottfried, H., & Yang, S. (2001). Forms/3: A first-order visual language to explore the boundaries of the spreadsheet paradigm. *Journal of Functional Programming, 11*(2), 155-206. https://doi.org/10.1017/S0956796800003828

Camilo, H., & Silva, J. A. (2008). OS testes de escolha múltipla (TEM) - The multiple choice tests (TEM). *Essências EDUcare*. Department of Medical Education, Faculty of Medicine-University of Coimbra. Retrieved from https://www.uc.pt/fmuc/gabineteeducacaomedica/fichaspedagogicas/Essencia_06

Čandrlić, S., Katić, M. A., & Dlab, M. H. (2014). Online vs. paper-based testing: A comparison of test results. In *Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 775-780). https://doi.org/10.1109/MIPRO.2014.6859649

Chan, N., & Kennedy, P. E. (2002). Are multiple-choice exams easier for economics students? A comparison of multiple-choice and "equivalent" constructed-response exam questions. *Southern Economic Journal*, *68*(4), 957-971. https://doi.org/10.2307/1061503

Clark, D. (2004). Testing programming skills with multiple choice questions. *Informatics in Education*, *3*(2), 161-178. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.149.3936&rep=rep1&type=pdf

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*, 1-17. https://doi.org/10.1016/j.ins.2012.10.039

Cortez, P., & Silva, A. (2008). *Using data mining to predict secondary school student performance*. Guimarães, Portugal: Universidade do Minho, Information Systems/ Algoritmi R&D Centre. Retrieved from https://repositorium.sdum.uminho.pt/bitstream/1822/8024/1/student.pdf

Costa, J. M., & Miranda, G. L. (2017). Desenvolvimento e validação de uma prova de avaliação das competências iniciais de programação [Development and validation of an assessment test of initial programming skills]. *RISTI – Revista Ibérica de Sistemas e Tecnologias de Informação*, *25*, 66-81. https://doi.org/10.17013/risti.25.66-81

Fernandes, N., Moro, S., Costa, C. J., & Aparício, M. (2019). Factors influencing charter flight departure delay. *Research in Transportation Business and Management*. In press. https://doi.org/10.1016/j.rtbm.2019.100413

Ferraz, A. P., & Belhot, R. V. (2010). Taxonomia de Bloom: Revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais [Bloom's taxonomy: Theoretical review and presentation of the adequacy of the instrument to define instructional objectives]. *Gestão & Produção*, *17*(2)*,* 421-431. https://doi.org/10.1590/S0104-530X2010000200015

Fuller, U., Johnson, C. G., Ahoniemi, T., Cukierman, D., Hernán-Losada, I., Jackova, J., ... Thompson, E. (2007). Developing a computer science specific learning taxonomy. *Proceedings of the 7th Annual Conference on Innovation and Technology in Computer Science Education* (pp. 152-170). https://doi.org/10.1145/1345443.1345438

Gronlund, N. E. (1988). *How to construct achievement tests*. Englewood Cliffs, NJ: Prentice-Hall.

Hickson, S., Reed, W. R., & Sander, N. (2010). To use constructed-response questions, or not to use constructed-response questions? That is the question. *Business and Law: Working Papers [69/2010]*. Christchurch, New Zealand: University of Canterbury. Retrieved from http://hdl.handle.net/10092/5418

Hudson, R. D. (2012). Is there a relationship between chemistry performance and question type, question content and gender? *Science Education International*, *23*(1), 56-83. Retrieved from https://files.eric.ed.gov/fulltext/EJ975550.pdf

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-mathematics items: It's not the strategy. *Journal of Educational Measurement*, *37*(1), 39-57. https://doi.org/10.1111/j.1745-3984.2000.tb01075.x

Kennedy, P., & Walstad, W. B. (1997). Combining multiple-choice and constructed-response test scores: An economist's view. *Applied Measurement in Education*, *10*(4), 359-375. https://doi.org/10.1207/s15324818ame1004_4

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory into Practice*, *41*(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Kuechler, W., & Simkin, M. G. (2003). How well do multiple choice tests evaluate student understanding in computer programming classes? *Journal of Information Systems Education*, *14*(4), 389-399. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.498.6912&rep=rep1&type=pdf

Lister, R. (2003). First year programming: Let all the flowers bloom. In *Proceedings of the 5th Australasian Conference on Computing Education* (pp. 221-230). Retrieved from https://pdfs.semanticscholar.org/c6ed/76cc2adf38370e914d90a6381982b0d24356.pdf

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, *31*(3), 234-250. https://doi.org/10.1111/j.1745-3984.1994.tb00445.x

Maresca, M. (2016). The spreadsheet space: Eliminating the boundaries of data cross-referencing. *Computer, 49*(9), 78-85. https://doi.org/10.1109/MC.2016.272

Moro, S., Esmerado, J., Ramos, P., & Alturas, B. (2019). Evaluating a guest satisfaction model through data mining. *International Journal of Contemporary Hospitality Management*. In press. https://doi.org/10.1108/IJCHM-03-2019-0280

Quintela, H. (2005). *Sistemas de conhecimento baseados em data mining: Aplicação à análise de estabilidade de estruturas metálicas* [Data mining-based knowledge systems: Application to stability analysis of steel structures].

Masters Dissertation. Braga, Portugal: Universidade do Minho. Retrieved from http://hdl.handle.net/1822/6255

Pinto, A. C. (2001). Factores relevantes na avaliação escolar por perguntas de escolha múltipla [Relevant factors in school assessment by multiple choice questions]. *Psicologia, Educação e Cultura, 5*(1), 23-44. Retrieved from https://repositorio-aberto.up.pt/bitstream/10216/18466/2/81873.pdf

Roberts, T. S. (2006). The use of multiple choice tests for formative and summative assessment. *Proceedings of the 8th Australasian Conference on Computing Education, volume 52* (pp. 175-180). Retrieved from https://dl.acm.org/citation.cfm?id=1151892

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*, 453 – 472. https://doi.org/10.1023/A:1003196224280

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85-106. https://doi.org/10.3102/1076998616666808

Silva, E. A. (2009). O ensino de estatística no curso de pedagogia usando o excel como instrumento facilitador da aprendizagem [Teaching statistics in pedagogy using excel as a learning facilitator]. *Revista Eletrônica Interdisciplinar*, 2. Retrieved December 2019 from: http://actividade.univar.edu.br/revista/downloads/estatistica.pdf

Silva, A. T., Moro, S., Rita, P., & Cortez, P. (2018). Unveiling the features of successful eBay smartphone sellers. *Journal of Retailing and Consumer Services, 43*, 311-324.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Science Journal of Innovative Education*, *3*(1), 73-98. https://doi.org/10.1111/j.1540-4609.2005.00053.x

Stankous, N. V. (2016). Constructive response vs. multiple-choice tests in math: American experience and discussion. In *Proceedings of the 2nd Pan-American Interdisciplinary Conference* (pp. 308-316). Retrieved from https://eujournal.org/index.php/esj/article/download/7410/7138

APPENDICES

APPENDIX A: EXCEL FUNCTIONS AND FORMULAS

TOPIC	FUNCTIONS EXAMPLES
Basic Functions	SUMIF; EXP; PRODUCT; SUMPRODUCT
Date and Time Function	DATE; DATEDIF; DAY; HOUR; TODAY
Formula	[Mix of Formulas and Functions]
Logical Function	IF; AND; FALSE; IFS; NOT; OR; TRUE
Search Function	VLOOKUP; ADDRESS; AREA; SELECT; COL; PROCH; INDEX; MATCH
Statistics Functions	COUNTIF; AVG; COUNTVAL; VAR; MAX/MIN
Text Functions	RIGHT; CONCATENATE; CODIGO; CONCAT; EXACT; SEG.TEXT

APPENDIX B: FEATURES DESCRIPTION

NAME	TYPE	DESCRIPTION
AnsweredCR	Logical	If the student answered the CR question or not
Topic	Factor (7)	Discipline Topic based in Excel Formulas
Difficulty	Factor (3)	{Easy, Medium, Hard}
BloomLevelMCQ/CR	Factor (4)	Bloom´s Taxonomy Levels
Gender	Factor (2)	{M/F}
Course	Factor (5)	{CE, CSBM, IC, …}
ExamPeriod	Factor (2)	Period when the student made the exam
Schedule	Factor (2)	{Daytime/ Evening}
NrSimilar	Factor (4)	Number of choices similar to the correct answer
NrDistractors	Factor (4)	Number of choices considered distractors to the correct answer
NrWordTextMCQ/CR	Numeric	Number of words of the question
NrCharacter-TextMCQ/CR	Numeric	Number of characters of the question
ScoreDifference	Numeric	Difference between CR and MCQ marks

BIOGRAPHIES

Yolanda Belo is Support Manager for Customer Loyalty and Customer-Relationship Management at GALP. She holds a M.Sc. in Computer and Business Management from ISCTE-IUL. Her research interests include data analysis to educational datasets.

Sérgio Moro, Ph.D. and Habilitation in Information Sciences and Technologies, is an Assistant Professor at ISCTE-IUL and Coordinator of the Information Systems Group at ISTAR-IUL. Sérgio is an interdisciplinary Data Scientist aiming to unveil patterns of knowledge through data-driven approaches in real real-world problems. His research appears in journals such as Decision Support Systems, Annals of Tourism Research, International Journal of Information Management, Expert Systems with Applications, Journal of Hospitality & Tourism Research, Computers in Industry, and Journal of Information Science, among others.

António Martins, PhD Business Administration, in the specialty of Strategy and Entrepreneurship, is Professor at ISLA Leiria, research of the Information Systems Group at ISTAR-IUL. Scientific Areas: Management informatics: management information systems, information systems management, project management and planning; Technological and organizational innovation.

Pedro Ramos is Associate Professor at ISCTE-IUL, and Researcher at both IT-IUL and ISTAR-IUL. He holds a PhD in Science and Information Technology, Master in Management Information Systems and a BSc In Management. He is the scientific coordinator of the Information Systems research field at DCTI/ISCTE. Coordinates several IT courses at ISCTE-IUL. Has large experience in the development of computer applications to the industry.

Joana Martinho Costa is Invited Assistant Professor at ISCTE-IUL. Her academic background includes a PhD in Education in the specialty of Information and Communication Technologies, a Master's in Computer Education and a BSc in Computer Engineering. She has published their works in several journals in the domains educational technology and Computer Science, namely the British Journal of Educational Technology, Informatics in Education, Informatica and Iberian Journal of Information Systems and Technologies. Her research interests are related to teaching methods applied to computer science, technology enhanced learning and statistical analysis.

Joaquim Esmerado (joaquim.esmerado@iscte-iul.pt, PhD in Graphical Computation from the University of Lausanne, Switzerland) is an Assistant Professor at Instituto Universitário de Lisboa (ISCTE-IUL), Portugal, and member of ISTAR-IUL. His scientific research interests include graphical computation as well as data and visual analytics applied to real world problems. He has published in journals such as International Journal of Information Management, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Computer Graphics and Applications, and International Journal of Advances in Computer Science and Its Applications.