Unfolding the drivers of students' success in answering multiple-choice questions about Microsoft Excel

1092

Many university programs include Microsoft Excel courses given its value as a scientific and technical tool. However, evaluating what was effectively been learned by students is a challenging task. Considering multiple-choice written exams are a standard evaluation format, this study aims at uncovering the features influencing students' success in answering this type of questions. The empirical experiments were based on Excel evaluation exams containing questions answered by 526 students between 2012 and 2016, in a total of 3,340 answers characterized by 17 features. Through data mining, a neural network was trained that accurately models students' choices. A sensitivity analysis was applied to the model to assess the most relevant features. Findings identified four highly relevant features for students' success: number of words of the question, topic, difficulty degree, and number of similar choices., This study helps to guide the design of future exams by quantifying the individual influence of each feature.

Keywords: multiple-choice questions; students' performance; Excel; data mining; feature relevance

Introduction

Spreadsheet tools are among the most used software applications worldwide (Tyszkiewicz, 2010). Particularly, Microsoft Excel© (from here forth, just Excel) is the most well-known spreadsheet application. Excel is one of the most powerful Microsoft brands, as users commonly refer to Excel as a synonym for spreadsheet or even tabular format (Barreto, 2015). As such, many bachelor programs worldwide teach how to use it, including in speciality areas such as Economics (Barreto, 2015) and medicine (Swallow, Newton & Van Lottum, 2003), among others, in focused courses, enabling university students to learn how to make the most of this powerful tool that will be used in the remaining courses of their degrees and also in their future professional lives. Although Excel is designed as a user-friendly tool, it is still difficult for students to learn how to use the more advanced features and for lecturers to effectively evaluate the knowledge students have acquired (Sitzmann, Ely, Bell & Bauer, 2010). Therefore, it is

important to find an adequate evaluation method that ensures the main topics were learned and to avoid students' enrolment in subsequent courses without adequate skills.

Written exams are the most adopted evaluation format worldwide. The process is simple: the student needs to answer questions correctly to show proficiency in the topic, with the exam being graded with a numerical or ordinal score. A grade above a predefined threshold (in most cases, 50% of the total score) grants students to pass the course. Thus, examiners are faced with the challenge of developing questions that accurately assess students' proficiency.

There are several types of questions. The two types mostly adopted are: multiple-choice questions, where a few choices are presented and students selects the correct one; and open-ended questions, where students develop a written answer at their will (Dochy, Janssens & Struyven, 2005). Literature has identified both advantages and disadvantages of using each type (Scouller, 1998; Dochy et al., 2005; Brown, Bull & Pendlebury, 2013). Multiple-choice questions pose serious challenges during their design due to their complexity and fast depletion which requires a large number of available questions (Pinto, 2001). Unlike multiple choice questions, open questions are faster and easier to build, being appropriate for analyzing the creativity and the reasoning of the student (Costa & Miranda, 2017; Pinto, 2001). However, this type of questions is harder to evaluate due to the subjectivity of human language expressed in written answers. Due to the latter fact, massive e-learning courses typically adopt multiple-choice questions (Zlatović, Balaban & Kermek, 2015). As a result of each type's strengths and weaknesses, several teachers choose to write exams using both types (McCoubrie, 2004). The present study focused specifically on multiple-choice questions used to evaluate students from Excel courses taught between 2012 and 2016. The main goal is to understand what drives the success in students' scores on multiple-choice questions.

The empirical research consisted of building a model with a dataset composed of 3,340 answers to multiple-choice questions covering Excel formulas and functions, from 526 different students, including several characterizing features such as question length and number of similar distractors to the correct one to assess which of those features played a more significant role in students' success. The model was built through data mining techniques, a procedure that was also adopted by Cortez and Silva (2008) to assess High School students' performance. The contributions of this study can guide the design of future multiple-choice questions about Excel.

Background

Educational data mining

Information systems have developed over the years by benefiting from an increase in processing performance as well as in storage capacity. Currently, databases store data about any business and are key for decision support as these conceal historical information which can be used for learning from the past to prepare for the future (Sharda et al., 2013). Specifically, nowadays educational institutions rely on information systems to manage their daily operations. Thus, their databases store information about applicants, students, courses and programs, which may be explored to answer important educational questions (e.g., what is the student success rate?). In the 1990s, Business Intelligence emerged to encompass all techniques and tools aiming to provide managerial intelligence through data analysis (Sharda et al., 2013). Within the Business Intelligence umbrella, data mining includes techniques and methods to uncover patterns from raw data that translate interesting findings which help in decision support (Moro et al., 2014). The main advantage of data mining in comparison to traditional statistical techniques is that artificial intelligence, which is included in data mining, enhances data modeling through complex

techniques that are able to apprehend non-linear relations between the features that characterize a given problem. In data mining, problems can be divided into supervised learning and unsupervised learning. In the former, a target feature translates the problem (e.g., the student's score in a course), while in the latter, there is only a set of input features and the goal is to find interesting relations among those features based on input data. For supervised learning, if the target feature is numerical, then it is a regression problem; otherwise, it is a classification problem (Sathya & Abraham, 2013).

Grounded on both data mining and educational information systems, educational data mining is a sub-field of data mining focused specifically on educational problems that can be addressed using data (Romero & Ventura, 2010). Several objectives may guide educational data mining projects, including predict students' results in courses, select weak students to propose additional classes, recommend to learners' activities to improve the learning experience, and evaluate the structure of the course content and its effectiveness on the learning process (Romero & Ventura, 2007). Regarding the prediction of students' score in a test, it is a supervised learning that can be considered a regression problem, if the score is numerical. However, in multiple-choice questions where the student needs to choose one among several possibilities, there are only 3 possible outcomes: correct answer, incorrect answer (with penalty), non-answered (no penalty). Therefore, it becomes a classification problem.

Evaluation through multiple-choice questions

The use of multiple-choice questions to evaluate the topics learned by students was introduced in the early 1900s in the United States as part of an educational reform to disseminate education throughout the entire country (Resnick & Resnick, 1992). The goal was to use an

objective and evaluator-independent method to evaluate student learning, thus eliminating the subjectivity inherent of open questions. Since then, this simple and effective method of assessing students' knowledge acquired during courses has spread worldwide. Several formats of choice questions are currently used, with the most popular including multiple-choice questions, where the student needs to select the correct choices (Roediger III & Marsh, 2005), and questions with independent sentences where the students need to signal if each sentence is true or false (Tsai, Tsai, Chai, Sung, Doong & Fung, 2004) or pairing questions (Pinto, 2001). Designing proper multiple-choice questions is a challenging task. Boland et al. (2010) identified two fundamental properties that exam writers need to be aware of: discrimination, to assure only knowledgeable students can find the correct choice; and the difficulty, which measures the probability of students knowing the topic but failing to find the correct choice. Both properties are complementary, meaning that writing the question requires a balance between these two properties. In multiple-choice questions where the student needs to select a single correct item among a list of them, the incorrect items are often called distractors (Boland, Lester & Williams, 2010). The difficulty in answering such questions is also associated to the plausibility of distractors, with a larger number of similar items to the correct one increasing the difficulty (Tarrant & Ware, 2008; Pinto, 2001). On the other hand, an easily rejected distractor will reduce the probability of getting the answer by exclusion (Pinto, 2001). One method often employed to discourage students' guessing when they do not know the topic addressed by the question is using penalties (Burton, 2001). Thus, students who are unaware of the topic are discouraged to answer, since failure is penalized with a negative score. However, "scores may be affected by the excessive reluctance of some examinees to risk answers" (Burton, 2001, p. 42). It is also important for each question to be based on a unique idea to be assessed and the vocabulary used

should be accessible to the students. Otherwise, the question evaluates the reading capability and not the relevant knowledge (Pinto, 2001). Although apparently multiple-choice questions are simpler and therefore preferred by students when they have no penalty for wrong answers (Pinto, 2001), their complexity can be quite large whether the distractors are very similar to each other or the number of choices is large. The questions should also be peer-reviewed to analyze possible interpretations and reduce ambiguity (Costa & Miranda, 2017; Miranda, 1985). The correction system for this type of questions is quite reliable and correction errors are usually insignificant (Pinto, 2001). Given the relevance of multiple-choice questions as an evaluation method, education researchers have devoted attention to understanding the main drivers of students' success in answering this type of question. The length of the question has been pointed out as an influencing factor of success, since lengthier questions typically require more effort in understanding them (Stiller et al., 2016; Yan & Tourangeau, 2008). Another known factor influencing difficulty is supporting images, which can help to reduce the difficulty of understanding the question (Stiller et al., 2016). Thus, existing literature highlights several features influencing success. Nevertheless, no study was found adopting a data mining approach to model student's success in answering multiple-choice questions. It should be noted that data mining sets its roots on both statistics and artificial intelligence, benefitting from both sciences to enable building non-linear complex models that can reveal patterns of knowledge (Moro, Cortez & Rita, 2014). Therefore, this study uses advanced computational modeling to provide guidance on creating multiple-choice questions applied to Excel, by using data mining, which has already been applied in other educational contexts (e.g., Monk, 2005; Cortez & Silva, 2008).

Microsoft Excel

Spreadsheet tools enable users to structure their information in a tabular format, thus organized in indexed rows and columns that provide a means to univocally identify a cell. Information is inserted or edited within each cell, with several types of data being allowed (e.g., numerical, text, dates). Also, formulas can combine the information from distinct cells, aggregate information, or use functions to compute new information based on an existing one. This type of tool emerged in the 1980s as a response to managerial and economical needs to take out the most of existing computational power to leverage decision making (Barreto, 2015). As a result, universities quickly began to include in their programs specific courses teaching how to use spreadsheet tools. Microsoft launched its Excel tool in response to this emerging and promising market at the time. During the 1990s, it quickly grew to dominate spreadsheet market, benefitting from integration with the remaining Microsoft Office tools, particularly, Microsoft Word (Barreto, 2015). Currently, it maintains the same status, although online tools such as the one provided by Google are gaining market share.

Excel is taught in a wide variety of programs, reflecting its multi-discipline nature. Accordingly, several researchers from different sciences have dedicated time and effort to understand and improve the learning process associated with Excel courses. Waldman and Ulema (2008) evaluated Excel teaching efficiency of an introductory information systems course by comparing the results of tests directly executed on the tool prior and after the course was taught. The same study considered tasks that used simple functions (e.g., logic functions such as IF), cell referencing, and cell formulas. Waldman and Ulema (2008) found that more than half of the students already had basic Excel skills, although the course helped them to improve those skills. Al Rawahi, Khan and Huq (2006) used Excel to integrate information technology into the mathematics curricula. As a result of their study, they advise teachers to develop integrated

methods using technology which, in turn, encourages students to have a more proactive attitude toward learning. Sitzmann et al. (2010) turned their attention to students' attrition in Excel online courses. The wide dissemination of Excel has earned Excel courses a spotlight place in e-learning, with online courses proliferating worldwide (Korn & Levitz, 2013). Still, no study was found aiming to uncover the influencing factors in Excel multiple-choice questions. Thus, the relevance of the present study is justified by both the widespread use of Excel and multiple-choice question exams, emphasized by the fact that typically most Massive Online Open Courses have an evaluation format based on multiple-choice questions (Margaryan, Bianco & Littlejohn, 2015).

Materials and Methods

Sample

The sample analyzed consists of hand-written Excel exams executed between 2012 and 2016 in a public university, located at Portugal, under two courses lectured, offered in several programs of the institution, mostly related to social sciences, technologies and information systems. The two courses are an introductory course on Excel, where the basic functionalities (e.g., spreadsheet formatting, simple arithmetic formulas) are taught, and an advanced course, which includes advanced spreadsheet formulas such as "lookup" tables and merging data. A total of 526 students were evaluated exclusively through 16 different exams. Each exam is composed of two sections, one with multiple-choice questions, and another with open-ended questions. The multiple-choice questions' section, which is the subject of the present study, ranges from 5 to 8 points score in a 20-grade scale (thus, from 25% to 40% of the overall grade). Considering the focus is on the answered questions, the dataset compiled consists in one line per question, in a total of 3,340 instances. The dataset encompasses a total of 16 different tests, with six different

questions per test. Each multiple-choice question accounts for 1-point score in a 20 scale and it has four choices, with only one being correct. Some of the remaining choices are implausible distractors, while others represent answers near the right answer. Since incorrect answers are penalized with a -0.25 score, a student may choose not to answer (0 score in that question). Thus, there are three possible categories for each multiple-choice question, correct, incorrect, not answered. Table 1 shows the distribution of scores for the analyzed questions.

<Table 1>

Table 2 shows all the features considered in this study. Since some exams may be overall harder than others, the exam identification itself was also included. The discipline topic was considered, as some topics may be harder than others (Lingard, Minasian-Batmanian, Vella, Cathers & Gonzalez, 2009). Table 3 displays the topic distribution, with formulas and basic functions accounting for more than half of the questions. In a first approach, to determinate the content value, as used in other studies (e.g. Costa & Miranda, 2017), all questions were validated by a specialist, a lecturer with more than 30 years of experience in teaching and evaluating Excel courses. However, since it is a subjective matter, the validation of each question was also performed by one of the co-authors and it was assured the agreement in all questions that were later answered by the students.. Also, since some of the questions were supported by images, a flag accounting for that was included. The degree of difficulty was assessed through the analysis of difficulty level following the Classical Test Theory resulting in a value between 0 and 1 for each question according to students' answers (Sartes & Souza-Formigoni, 2013). This value was later converted in qualitative measures: 'hard' for values below than 0.33, 'medium' for values between 0.33 and 0.66, and 'easy' for values above than 0.66.

<Table 2>

<Table 3>

Knowledge extraction

Extracting knowledge from raw data requires building a model that may reveal the hidden patterns that might be able to explain the influence of the input features identified in Table 2 when modelling the output (the answer grade). Several data mining models are adequate to build a classifier with the purpose of modelling the grade, including naïve Bayes (NB), decision trees (DT), support vector machines (SVM), and neural networks (NN) (Moro et al., 2014). NB assumes the independence of the input features between each other to build a "naïve" model (Rish, 2001), while DT defines a set of rules in a branch format where each branch denotes a decision based on one feature until a leaf of the tree is reached where the predicted outcome is shown (Loh, 2011). The most advanced machine learning techniques applied in data mining model complex relations between features to achieve more accurate models. These include both SVM and NN. The former defines separating hyperplanes using support vector points to distinguish data (Guerreiro & Moro, 2017), while the latter attempts to mimic human brain behavior by defining layers connecting neurons (or nodes), where their state is computed through a weight function using previous neurons' values (Haykin, 2009). The SVM transforms the complex input space into a high dimensional feature space through a nonlinear mapping that depends on a kernel. There are several types of NN. We adopted the popular multilayer perceptron, configured with one hidden layer composed of a set of hidden nodes configured through a grid search under the conditions advised by Moro et al. (2014). The NN's nodes apply a logistic function with the input values from the connections to compute an output. Thus, the

NN complexity derives from the number of nodes and their connections. The four aforementioned modelling techniques were tested, with its results evaluated using two metrics: the area under the Receiver Operating Characteristic (ROC) curve (AUC); and the area under the cumulative Lift curve (ALIFT). The ROC curve plots the false-positive rate versus the true-positive rate, with a random classifier being represented by a diagonal from 0 to 1. Thus, random and optimal classifiers hold an AUC value of 0.5 and 1.0, respectively. The cumulative Lift curve is plotted with a set of instances sorted from the most to the least likely target outcome (i.e., a correct answer if we are trying to test the accuracy of the model in predicting correct answers). Therefore, a better classifier will achieve an ALIFT closer to 1.0. Both classification metrics were used and described in detail by Moro, Cortez and Rita (2015).

All the experiments were conducted using the R statistical tool and the "rminer" package, specifically designed to facilitate data mining experiments (Cortez, 2010). To further assure modelling robustness, a 10-fold cross-validation scheme was adopted (Moro, Rita & Coelho, 2017b). The results for the four techniques tested are shown in Table 4. There are several metrics to assess a classifier's performance, i.e., a model in which the target feature is a range of classes (in our case, "Incorrect", "Correct", and "Non-answered"). We adopted AUC and ALIFT, which were previously described. For both cases, a random classifier is represented by the 0.5 value, while an ideal classifier (i.e., one that accurately classifies every cases) holds the value of 1.0. Thus, the higher the values shown in Table 4, the better is the classifier.

<Table 4>

Table 4 highlights with no surprise that the most advanced techniques of SVM and NN outperform the remaining when classifying the three possible classes for a multiple-choice

question grade. Furthermore, NN reached the best values of both AUC and ALIFT, a result aligned with the study by Moro et al. (2014). Therefore, the NN model was selected for knowledge extraction.

NN provide an accurate model, although directly understanding it is beyond human comprehension, since the rules that led to it are not directly obtained (Browne, Hudson, Whitley, Ford, Picton & Kazemian, 2003). However, there are a few effective techniques for extracting knowledge from NN, namely rules extraction and sensitivity analysis. The former has the disadvantage of simplifying the inherent complexity of NN in the attempt to extract rules, thus, losing information, while the latter enables one to assess the outcome variation when changing the input features, offering an interesting approach to unveil features' contribution in modelling the outcome (Cortez & Embrechts, 2011). The data-based sensitivity analysis (DSA) is a sensitivity analysis technique which uses a randomly selected sample of data to test model variation by simultaneously changing the input features, thus being able to disentangle the relationships between features (Cortez & Embrechts, 2013). DSA has been successfully applied to a wide variety of problems, such as tourism (Guerreiro & Moro, 2017), social media (Moro, Rita, & Vala, 2016), groundwater (Naghibi, Pourghasemi & Abbaspour, 2017), and banking (Moro, Cortez & Rita, 2017a); thus, it was chosen for the present study, in order to understand how each of the input features contribute to the most accurate model trained, i.e., the NN.

Results and discussion

The DSA enabled us to obtain the relative relevance of each feature to modelling the type of answer (Correct, Incorrect, and Non-answered) to multiple-choice questions. Figure 1 shows how each input feature described in Table 2 contributed to the model. The length of the question measured by the number of words was considered the most relevant feature. This result is

aligned with the findings reported by Stiller et al. (2016), who also discovered a significant effect of the word count on a 63 multiple-choice questions test answered by 907 participants to assess the pre-service teachers' scientific reasoning abilities. Furthermore, it is interesting to highlight that several studies on distinct evaluation subjects also found similar results, such as Martiniello (2009) did on mathematic tests. A similar effect is known from surveys, with longer questions resulting in higher response times (Yan & Tourangeau, 2008). Regarding the number of words in each question and the difficulty degree of our sample, the results showed that, for each average of words (Easy = 17.3; Medium = 19.7; Hard = 16.8), there were no evidence of significant correlation (Spearman's $\rho = 0.093$). Thus, the results presented are consistent with existing literature. Freedle and Kostin (1996) studied the association between the length of a sentence and its difficulty degree and they found no significant effects for listening words. Nevertheless, none of the aforementioned authors analyzed the impact of word count in students' choice of not answering the question, as it requires case-studies of questions where a penalty (negative score) is attributed to selecting an incorrect choice (Burton, 2001).

<Figure 1>

The second most relevant feature represents the different Excel topics that were evaluated. While there are several studies analyzing students' success in learning Excel (e.g., Bai, 2009; Frydenberg, 2013), none was found categorizing the different topics covered. Nevertheless, this is an expected result for any type of question. A third finding appears to be related to the difficulty degree, which accounts for more than 12% of relevance in students' success rate. Lingard et al. (2009) found that student and staff agreement on perception of question difficulty is only about 50% for the disciplines analyzed, (biochemistry and physics).

On the other hand, as the same authors stated, different cognitive skills are required for the different sciences; on the opposite sense, Excel is a tool with application purposes, specifically designed to be user-friendly. Future research is needed to scrutinize what is a difficult question for the case analysed. The number of similar choices to the correct answer emerges in fourth place, with almost 10% of relevance accounted for. This is a known and intuitive effect, also reported in the literature (Burton, 2005), as more similar choice items imply fewer implausible distractors, thus students are more likely to miss the correct answer in the presence of a larger number of similar choices to the correct one. The remaining features hold individual relevance of around 8% or less.

Next, the four most relevant features are scrutinized, all of them with an individual relevance of around 10% or above. While there are three possible outcomes (Correct, Incorrect, and Non-answered), an incorrect answer is the opposite of a correct answer. Therefore, only the probability for the correct and non-answered outcomes are plotted and discussed.

Figure 2 shows the impact of question length in both correct or non-answered responses. The results are as expected: lengthier questions are harder to correctly answer, while simultaneously tend to lead students to avoid answering them. The contribution of Figure 2 lies in measuring such effect: from 10 to 30 words, the question rapidly increases in terms of difficulty measured by correct answers, while the effect is attenuated afterwards. Moreover, questions with a length from 20 to 35 words have a pronounced increasing probability of not being answered, until it reaches a plateau above that. Other studies lack in quantifying such effect (e.g., Stiller et al., 2016).

<Figure 2>

The second most relevant feature is the topic addressed by the question. Figure 3 displays the probability of choosing the correct answer or choosing not to answer a question depending on the six topics described in Table 3. There are some apparently surprising results, namely students are likely to correctly answer any text function question (93.1%), while basic functions hold a rather low probability for students choosing the correct choice (42.2%). Such finding calls for an analysis of two randomly chosen examples of questions from these two topics (Table 5). Both the chosen questions are "easy" (difficulty degree), with the text function question being longer in words. Such result raises the hypothesis that the lectures devoted more attention to text functions, assuming that basic functions such as ROUND were already known from university students. This discovery may help shape future lessons. However, the hardest topic is "referencing": not only students are having a hard time in correctly answering this type of question, but also it is the topic where students show less confidence, reflecting it in the highest number of non-answered questions, 5.7%. As Maresca (2016) pointed out, referencing can be used not only to interconnect information within the same spreadsheet, but also from different spreadsheets as well as external sources. Thus, it requires deep understanding about where the information resides, which may be difficult for bachelor students who are more used to the remaining mathematical concepts associated with functions and formulas offered by Excel.

<Table 5>

<Figure 3>

The results shown on Figure 4 for the difficulty degree, the third most relevant feature, are as expected: the probability of answering correctly is directly dependent degree of difficulty, with harder questions having a lower probability of being corrected answered. Also as expected,

harder questions are more likely to be left unanswered. Nevertheless, the effect of the difficulty degree is more pronounced for the case of correctly answering the question, implying that harder questions are not in some cases discouraging students to take their chance. While the findings are as expected, the match between the perceived difficulty by the attributed the difficulty degree and by the students (who answered or avoided answering the questions) is an accomplishment difficult to achieve (Linn, Baker & Dunbar, 1991; Lingard et al., 2009). As Dochy et al. (2005) pointed out, it is important to understand students' perceptions during the learning process that will help them succeed in the evaluation. Nevertheless, while some disciplines such as physics require a complex rational to answer each question (Lingard et al., 2009), Excel's practical perspective may help to explain why there is an alignment between the perceived difficulty degree of the lecturer and students' performance.

<Figure 4>

Lastly, the number of similar choices to the correct answer is scrutinized through Figure 5. As a higher number of similar choices represents a lower number of implausible distractors (since the two are complementary), the more similar choices, the more likely the student fails in correctly answering the question. The opposite happens for the case of not answering the question, since a larger number of similar questions can raise doubts in students who may prefer not to answer. This is a known effect from the literature, as writing multiple-choice questions requires to balance the number of possible choices between plausible (similar) and implausible items (Boland et al., 2010). This study helps by quantifying such effect for the case of Excel, including multiple-choice questions with penalty for an incorrect answer.

<Figure 5>

Conclusions

Microsoft Excel is a valuable tool recognized and used in a myriad of roles in the labor market worldwide. Therefore, it is taught in many undergraduate programs in most universities. As with any subject taught in a university course, it requires an evaluation process that is adequate for assessing if students learned the different topics taught. The present study analyzed the case of a curricular unit lectured between 2012 and 2016 at a European university where tests included both choice and development questions. Particularly, a model aiming to accurately model students' answers categorized in three types (correct, incorrect, and non-answered) was built from where knowledge was extracted to assess features contribution to modelling each of the types.

Results enabled to us to identify four features holding around half of the relevance to students' score, from the total of 17: (1) number of words of the question, (2) topic, (3) difficulty degree, and (4) the number of similar choices. The effect of each of these four features was assessed. Lengthier questions are harder to answer, and the students are more likely to choose not answer those. The topic of "referencing" was found to be difficult and particularly discouraging for students who, in some cases, chose not to answer the corresponding questions.

The major contributions include quantifying the causal effect of known features on the possible outcomes of choice questions where an incorrect choice receives a penalty, discouraging students' uninformed guesses, and unveiling how each feature is affecting students' performance for the case of Excel. The findings may help to guide future lectures in designing tests and revising the subjects taught aiming at improving the process of learning to work with Excel.

Limitations and Future Work

Our study has some limitations that can be explored in future research. Firstly, we have analyzed 17 features (Table 3) that can influence the results of multiple-choice questions but mostly are related to the characteristics of the questions. In future work, we propose to understand whether the students' characteristics and academic background influence the success on multiple-choice questions since our research did not cover these factors. Our focus was also essentially related to the question. We intend to analyze the distractors more in depth since the work of Freedle and Kostin (1996) obtained similar results for listening words but they mentioned that the length of sentences in the distractors has a significance effect in the difficulty degree of the question. We propose to research this issue further.

The design was centered on Excel for its popularity among universities but there are a set of other spreadsheets that could also be researched. Our data suggests that the results will be similar in other spreadsheets, but it lacks on a confirmation due to the differences in the type of programming languages and formats between spreadsheets, which can influence the difficulty degree of the questions. We suggest confirming it in future research.

References

Al Rawahi, F. K., Khan, S. A., & Huq, A. (2006). *Microsoft EXCEL in the mathematics classroom: A case study*. In Proceedings of the 2nd Middle East Teachers of Science, Mathematics and Computing Conference (METSMaC), pp. 131-134.

Bai, H. (2009). Assigning Students in Cooperative and Individual Learning Environments According to Cognitive Styles: Achievement and Perceptions in Computer Technology Learning. *i-Manager's Journal on School Educational Technology, 5*(1), 7-16.

Barreto, H. (2015). Why Excel?. *The Journal of Economic Education*, 46(3), 300-309.

Boland, R. J., Lester, N. A., & Williams, E. (2010). Writing multiple-choice questions. *Academic Psychiatry, 34*(4), 310.

Brown, G. A., Bull, J., & Pendlebury, M. (2013). *Assessing student learning in higher education*. London: Routledge.

Browne, A., Hudson, B., Whitley, D., Ford, M., Picton, P., & Kazemian, H. (2003). *Knowledge extraction from neural networks*. In Industrial Electronics Society, 2003. IECON'03. The 29th Annual Conference of the IEEE (Vol. 2, pp. 1909-1913). IEEE.

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education, 26*(1), 41-50.

Burton, R. F. (2005). Multiple-choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education, 30*(1), 65-72.

Cortez, P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. *Advances in data mining. Applications and theoretical aspects*, 572-583.

Cortez, P., & Embrechts, M. J. (2011). *Opening black box data mining models using sensitivity analysis*. In Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on (pp. 341-348). IEEE.

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences, 225*, 1-17.

Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. In EUROSIS, A. Brito and J. Teixeira (Eds.), pp. 5-12.

Costa, J.M., & Miranda, G.L. (2017). Desenvolvimento e validação de uma prova de avaliação das competências iniciais de programação. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação, 25*, 66-81. Doi: http://dx.doi.org/10.17013/risti.25.66-81

Dochy, F., Janssens, S., & Struyven, K. (2005). Students' perceptions about evaluation and assessment in higher education: a review 1. *Assessment & Evaluation in Higher Education, 30*(4), 325-341.

Frydenberg, M. (2013). Flipping excel. *Information Systems Education Journal, 11*(1), 63-73.

Freedle, R., & Kostin, I. (1996). The prediction of TOEFL listening comprehension item difficulty for minitalk passages: Implications for construct validity. *ETS Research Report Series, 1996*(2), i-61.

Guerreiro, J., & Moro, S. (2017). Are Yelp's tips helpful in building influential consumers?. *Tourism Management Perspectives, 24*, 151-154.

Haykin, S. S. (2009). *Neural networks and learning machines* (Vol. 3). Upper Saddle River, NJ, USA: Pearson.

Korn, M., & Levitz, J. (2013). Online courses look for a business model. *The Wall Street Journal*, B8.

Lingard, J., Minasian-Batmanian, L., Vella, G., Cathers, I., & Gonzalez, C. (2009). Do students with well-aligned perceptions of question difficulty perform better?. *Assessment & Evaluation in Higher Education, 34*(6), 603-619.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14-23.

Maresca, M. (2016). The Spreadsheet Space: Eliminating the Boundaries of Data Cross-Referencing. *Computer*, *49*(9), 78-85.

Margaryan, A., Bianco, M., & Littlejohn, A. (2015). Instructional quality of massive open online courses (MOOCs). *Computers & Education, 80*, 77-83.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational assessment*, *14*(3-4), 160-179.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical teacher*, *26*(8), 709-712.

Miranda, M.J. (1985). Docimologia em Perspectiva. *Revista da Faculdade de Educação*, *8*(1), 39-69.

Monk, D. (2005). Using data mining for e-learning decision making. *The Electronic Journal of e-Learning*, *3*(1), 41-54.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22-31.

Moro, S., Cortez, P., & Rita, P. (2015). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. *Neural Computing and Applications*, *26*(1), 131-139.

Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, *69*(9), 3341-3351.

Moro, S., Cortez, P., & Rita, P. (2017a). A framework for increasing the value of predictive data-driven models by enriching problem domain characterization with novel features. *Neural Computing and Applications*, *28*(6), 1515-1523.

Moro, S., Rita, P., & Coelho, J. (2017b). Stripping customers' feedback on hotels through data mining: the case of Las Vegas Strip. *Tourism Management Perspectives*, *23*, 41-52.

Naghibi, S. A., Pourghasemi, H. R., & Abbaspour, K. (2017). A comparison between ten advanced and soft computing models for groundwater qanat potential assessment in Iran using R and GIS. *Theoretical and Applied Climatology*, DOI: https://dx.doi.org/10.1007/s00704-016-2022-4.

Pinto, A.C. (2001). Factores relevantes na avaliação escolar por perguntas de escolha múltipla. *Psicologia, Educação e Cultura, 5*(1), 23-44.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. *Changing assessments: Alternative views of aptitude, achievement, and instruction*, Springer.

Rish, I. (2001). *An empirical study of the naive Bayes classifier*. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). IBM.

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1), 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C* (Applications and Reviews), 40(6), 601-618.

Sartes, L., & Souza-Formigoni, M.L. (2013). Avanços na Psicometria: Da Teoria Clássica dos Testes à Teoria de Resposta ao Item [Advances in Psychometry: From Classical Theory of Tests to Item Response Theory]. *Psicologia: Reflexão e Crítica, 26*(2), 241-250. doi: http://dx.doi.org/10.1590/S0102-79722013000200004.

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2), 34-38.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. Higher Education, 35(4), 453-472.

Sharda, R., Delen, D., & Turban, E. (2013). *Business intelligence: a managerial perspective on analytics*. Prentice Hall Press.

Sitzmann, T., Ely, K., Bell, B. S., & Bauer, K. N. (2010). The effects of technical difficulties on learning and attrition during online training. *Journal of Experimental Psychology: Applied, 16*(3), 281-292.

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., ... & Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education, 41*(5), 721-732.

Swallow, V., Newton, J., & Van Lottum, C. (2003). How to manage and display qualitative data using 'Framework'and Microsoft® Excel. Journal of clinical nursing, 12(4), 610-612.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42*(2), 198-206.

Tsai, S. L., Tsai, W. W., Chai, S. K., Sung, W. H., Doong, J. L., & Fung, C. P. (2004). Evaluation of computer-assisted multimedia instruction in intravenous injection. *International Journal of Nursing Studies, 41*(2), 191-198.

Tyszkiewicz, J. (2010). *Spreadsheet as a relational database engine*. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (pp. 195-206). ACM.

Waldman, M., & Ulema, M. (2008). Automated measurement and analysis of effectiveness of teaching selected Excel topics in an introductory IS class. *Journal of Computing Sciences in Colleges, 23*(5), 73-82.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology, 22*(1), 51-68.

Zlatović, M., Balaban, I., & Kermek, D. (2015). Using online assessments to stimulate learning strategies and achievement of learning goals. Computers & Education, 91, 32-45.

Table 1 - Answers in the dataset.

Answer	**Number**
Incorrect	1,121
Non answered	219
Correct	2,000
Total	**3,340**

Table 2 - Features considered.

#	**Name**	**Description**
1	exam	Exam identification
2	scientific.area	{Sociology, Technology, Management}
3	course	{Advanced Excel, Introduction to Excel}
4	year	Year and month when the exam was made
5	month	
6	total.exams	Total number of exams made by every student in that year
7	topic	Discipline topic within 7 (see Table 3 for details)
8	difficulty.degree	{Easy, Medium, Hard}
9	nr.sim.choices	Number of choices similar to the correct answer
10	image	If the question was supported by an image
11	choice.quest.grade	Grade and weight of the multiple-choice questions' section, measured in a value where 20 represents the maximum possible grade in the exam
12	choice.quest.weight	
13	open.quest.grade	Grade in the essay questions' section
14	exam.grade	Total grade in the exam (in a 20 grade scale)
15	quest.type	{Theory, Application}
16	quest.nr.words	Number of words of the question
17	quest.nr.chars	Number of characters of the question
18	answer.grade	Grade achieved by the student with his/her answer (Table 1)

Table 3 - Question topics.

Topic	**Example**	**Total questions**
Logic functions	IF function	585
Referencing	write function to fill in a specific cell	75
Text functions	MID function	78
Basic functions	ROUND function	878
Formulas	find the formula to match the sheet shown	1,123
About Excel	shortcut keys	601
	Total	**3,340**

Table 4 - Modelling results.

Model	Correct		Incorrect		Non-answered	
	AUC	**ALIFT**	**AUC**	**ALIFT**	**AUC**	**ALIFT**
NB	0.785	0.615	0.749	0.666	0.773	0.754
DT	0.751	0.601	0.715	0.643	0.706	0.692
SVM	0.828	0.633	0.794	0.695	0.803	0.783
NN	0.830	0.633	0.795	0.696	0.816	0.795

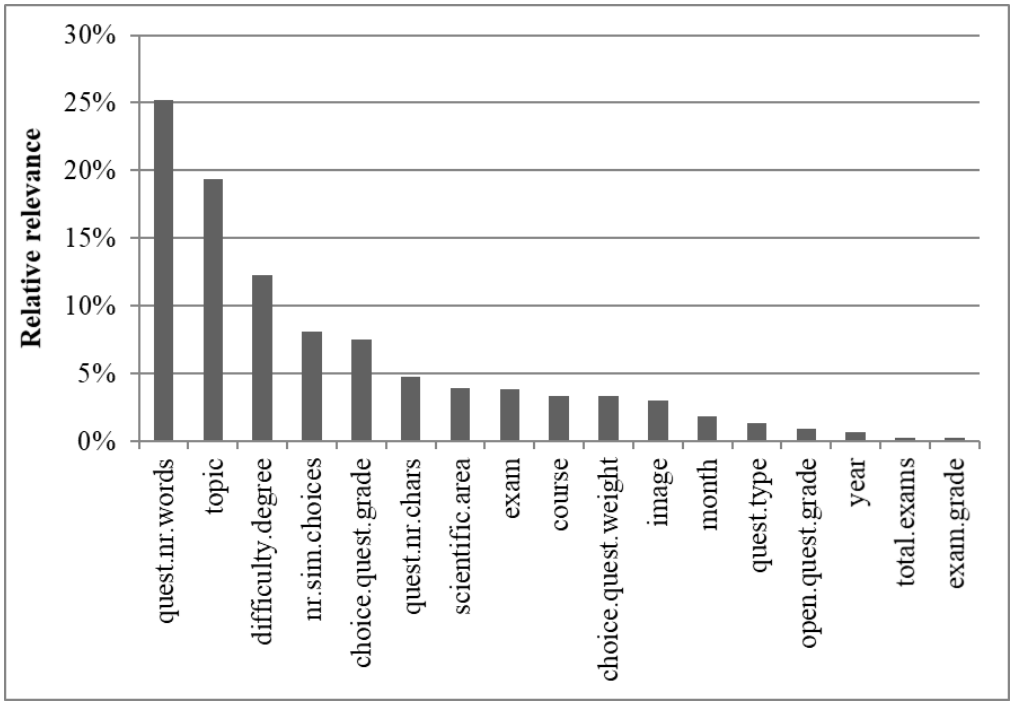

Figure 1 - Relative relevance of each feature to the model.

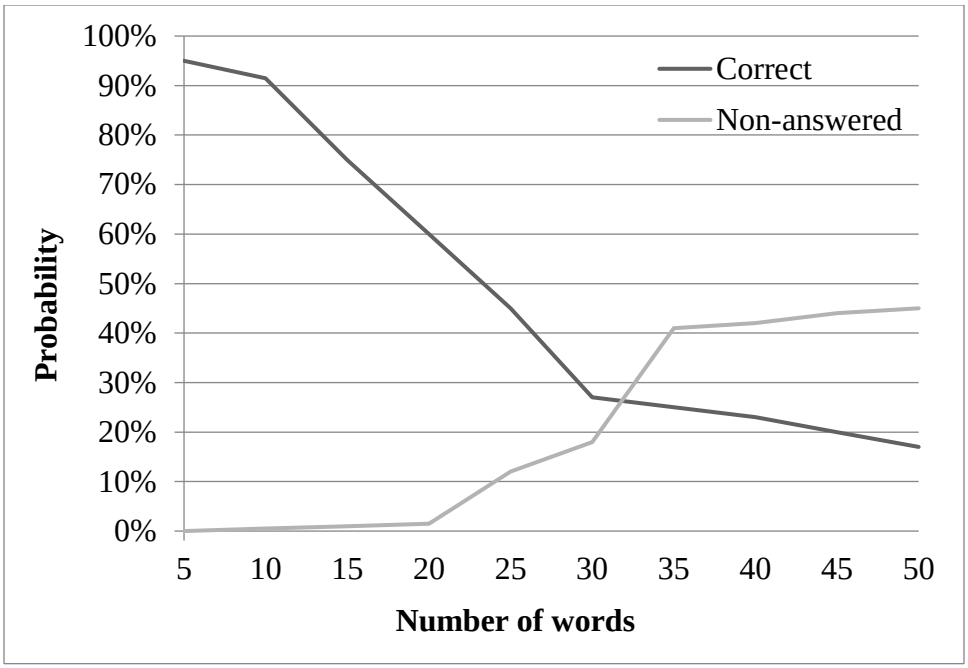

Figure 2 - Probability for the type of answer depending on the number of words.

Table 5 - Examples of text and basic functions.

Topic	**Question**	Choices (in **bold** the correct one)							
Text function question	Consider it is written in cell A1 "Excel function". If, in A2, you insert MID(A1,3,2), the result is:	**1**	**"ce"**	2	"ex"	3	"cti"	4	"xce"
Basic function	What is the result of the following function: round(128.45,-1)	1	120	2	128	**3**	**130**	4	131

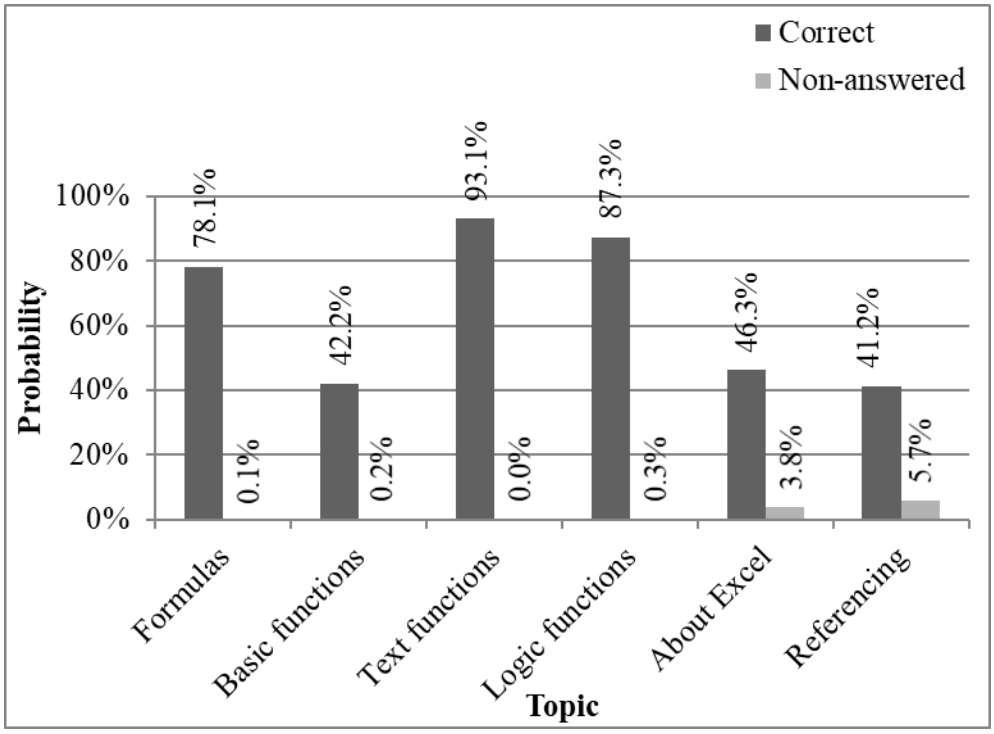

Figure 3 - Probability for the type of answer depending on the topic.

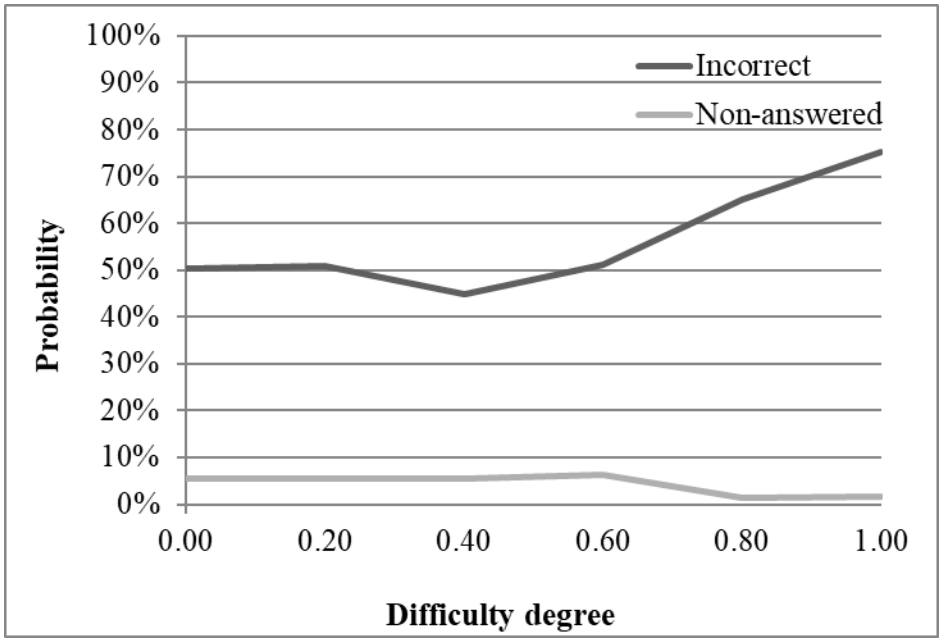

Figure 4 - Probability for the type of answer depending on the difficulty degree.

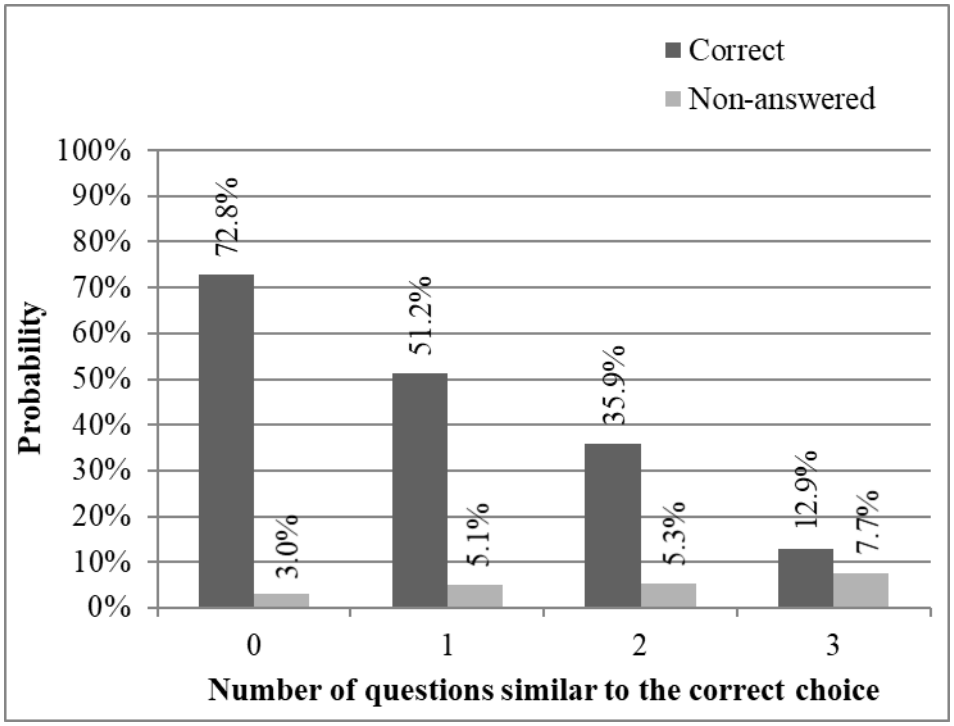

Figure 5 - Probability for the type of answer depending on the number of similar questions.