A methodology for early detection of potentially low-achievement students by means of on-line testing and Item Response Theory

D Blanco¹, F Sánchez¹, B J Álvarez¹, P Fernández¹* and N Beltrán¹

¹ Department of Construction and Manufacturing Engineering, University of Oviedo

*Corresponding author: pedrofa@uniovi.es

Abstract: There is an increasing concern about the relevance of reducing university dropout, since the available information shows how dropout reaches surprisingly high values. Additionally, the particular dynamics of university courses makes it difficult to identify students that are prone to fail the course. Present work proposes a methodology to implement an early detection of students in risk of failure, based on the Item Response Theory (IRT) analysis of data collected from on-line tests. Questions are presented to the students according to knowledge campaigns within the frame of the course by means of an on-line knowledge management platform (ZAPIENS). Results from student’s performance are processed using IRT to obtain the value of ability for each individual. This information is later used to identify those students that, with high probability, would fail the subject. Students would be grouped according to a Homogeneous Cluster Strategy into different ability levels. Accordingly, faculty would be capable of implementing specific reinforcement strategies. Results indicate that the methodology allows for an appropriate segmentation of the students according to their capability at an early stage. This strategy could help in the development of differential learning itineraries and could also be easily adapted to different educational environments.

Keywords: Evaluation methodologies, Data science applications in education, Architectures for educational technology system.

1. Introduction

The European Commission has included an objective of 40% of the younger generation (individuals aged 30-34) having a tertiary degree among the goals of the Europe 2020 strategy. This would imply a 9% increase in just a decade. This objective reflects the importance that success in higher education has on the economy of the EU and its knowledge-intensive sectors. Reducing dropout turns to be, consequently, a key factor in the educational strategy of the Union [1]. In Spain [2], the dropout after the first year (2012/13) reaches a 20.25%, while only 49.26% of the students obtain their degree in the expected time (data from four-year degrees analysis in course 2016/17). The results are even worse in the field of Engineering and Architecture, presenting a 32.43% first-year dropout and a mere 31.28% of students obtaining their degree in four years.

In the Polytechnic School of Engineering at Gijón, there is a subject called “Manufacturing Processes”. This subject provides introductory-level knowledge concerning manufacturing engineering. Traditionally, the subject scores under average regarding main academic parameters. The evaluation rate, defined as the percentage relationship between the number of students presented to evaluation and
the total number of students enrolled, has reached a mere 67.1% in the 2016-2017 course. This means that a 32.9% of the enrolled students do not even present themselves to evaluation. Moreover, this problem is not exclusive of Manufacturing Processes, since the overall dropout rate (DR) for some degrees even tripled the expected results (60.8% DR was scored in 2013-2014 within the Mechanical Engineering degree). The fact that only one out of every four engineering students in Asturias manages to finish the correspondent degree in the foreseen time, indicates that there is a serious problem of inefficiency that must be addressed [2].

Dropout is influenced by many factors that should not be analyzed individually, since they are interconnected. Possible approaches to reduce high dropout rates can focus on acting upon the socio-economical causes [3,4] or on providing individual support to those students prone to dropout. Although both types of approaches should be implemented simultaneously, the act-upon-causes approach is related to educational and even social-politics and falls out of faculty’s range of action. On the other side, the individual-support approach could only be implemented if an adequate tool for identifying students in risk of failure has been made available.

In recent years, the development of on-line tools and resources, alongside with the generalization of connectivity through mobile devices (smartphones, laptops and tablets) in the educational environment, has helped many researchers to implement different methods for early warning [5]. Some works try to predict the performance of students [6] whereas others try to identify students with a high risk of failure by predicting if they would pass or fail a course [7]. The parameters used in these researches came from the performance of the students in early tasks or intermediate tests and exercise grades [8]. The mathematical methods and computational approaches used for prediction also cover a wide range of techniques, from linear regression [5] to fuzzy logic [7].

For this work the item response theory (IRT) or latent trait theory [8, 9] has been used for linking the ability of an individual with the score obtained in a test. Using IRT, students that answer a test are ranked according to the latent trait (i.e.: their “ability” or capability for providing the right answer), whereas each question or item is characterized according to its capability to discriminate those students.

Data collected from the test will be processed by means of IRT, in order to obtain individual values for the ability (hidden trait) of the students that could be used for segmentation of the students into different levels of failure risk. A proof of concept of the proposed methodology is also presented in this work. Therefore, ability levels have been calculated for engineering students enrolled in Manufacturing Processes courses at EPIG during the 2017-18 academic year and later used for clustering purposes. The accuracy of this clustering strategy would be finally tested upon the 2018-19 students.

2. Experimental Planning

2.1. Item Response Theory

IRT provides several alternative models that could be used to calculate the probability of a particular individual answering correctly a certain question. In this work, the two-parameter model has been used. This model is capable of establishing a value for both the difficulty of the question and its corresponding level of discrimination. The difficulty reflects the percentage of students that answer the question correctly. The discrimination reflects the capability of the question to differentiate among the students on the basis of its knowledge of the subject being tested. IRT assumes the existence of a functional relationship between the values of the variable (θ) that reflects the level of knowledge (also known as ability) of an individual and the probability of him giving the right answer to a particular question.

The P-function represents the probability of the i-th individual (with a θi level of knowledge) answering correctly to the j-th question (with a discrimination level ai and a difficulty level bi). In this work, the scaling constant 1.702 is chosen so that the absolute difference between the normal and logistic cumulative distribution function is less than 0.01 over the real line [10]. The P-function [10, 11], can be expressed as (1):

\[
P(u_j = 1|\theta_i, a_j, b_j) = \frac{\exp [1.702 a_j (\theta_i - b_j)]}{1 + \exp [1.702 a_j (\theta_i - b_j)]}
\] (1)
Given a group of individuals that perform a series of tests, the IRT would provide a parameter that reflects their ability and then generate curves that reflect the probability of answering correctly each individual question as a function of that ability. Therefore, the likelihood of this probability depends on both the difficulty of the question and the ability of the students. The calculation of the referred parameters is performed by means of the statistical software and programming language R [10].

2.2. Methodology description

The main objective of this work is to determine if the ability parameter calculated from a series of tests could be used for discriminating students according to an estimation of how likely they would pass or fail the subject. Accordingly, once the individual ability values have been calculated, the goal would be to determine if there is a significant relationship between the level of ability and the likelihood of the student passing the course. Therefore, a discriminating value of ability – High Risk of Failure limit (θ_{HRF}) – would be calculated, so that students that would score ability values under this limit would have high possibilities of failing the course. These students would be clustered in a “low ability group” (LAG). In a similar way, a discriminating value of ability – High Probability of Passing the course limit (θ_{HPP}) – should be calculated, so that students showing higher abilities would have the highest possibilities of passing the course, and they would be clustered in a “high ability group” (HAG). The rest of the students, with intermediate levels of ability, would integrate the “medium ability group” (MAG).

Two alternative approaches could be taken for establishing the limits: fixing the percentage of students that would be assigned to LAG and HAG a priori, or fixing the minimum percentage of correct assignment acceptable within LAG and HAG. The researchers have discussed this problem and finally decided to adopt a balanced group criterion, with leads to have 20% of students assigned to the LAG, 20% assigned to the HAG and 60% of students clustered in the MAG. These percentages would be used to determine θ_{HRF} and θ_{HPP} during the 2017-18 academic year, so that both levels would be used during the 2018-19 course to check the accuracy of HCS. The comparison between the ability results obtained by the students during the test and their final results (pass/fail) could only be done once the semester is over.

Additionally, since an “early” detection of students with high risk of failing the course should be the ideal outcome of this research, it should be determined if an adequate clustering of students does necessarily require of an intensive multi-theme set of test or if it could be also possible based on a less time-consuming single-theme set of questions. To fulfill all these objectives and limitations, the methodology reflected on figure 1 was adopted.

Accordingly, the whole course should be organized and divided into “BLOCKS” of homogeneous knowledge contents. Each block of knowledge would correspond to a so-called campaign that would comprise a certain number of individual test questions. Consequently, during the whole semester, students’ progression would be monitored by means of test questions. The result of this process is a value for the ability of each single student with respect to the ability of the group within the limits of the correspondent campaign. The sequence follows through all the knowledge blocks and, once the semester has finished and the last campaign concluded, a global analysis of student’s ability must be additionally performed by means of IRT.

The final exam evaluates students’ knowledge, and their final mark would be calculated. Then, results from the exam (success/failure) and results from the IRT calculation (ability level) would be analyzed in order to find out if there is a correlation between both results. This analysis would be performed for each individual campaign and for the whole subject. This should lead to θ_{HRF} and θ_{HPP} calculations, which would allow the students to be clustered into the three groups previously described. Finally, during the following course the procedure should be repeated, but this time, the new cohort of students would be clustered according to θ_{HRF} and θ_{HPP} values, as they were calculated during the previous semester. When the final results would be available for those students, it would be possible to quantify the accuracy of the clustering strategy and thus its capability for early clustering of students.
2.3. Implementation and campaign structure

ZAPPIENS platform, although initially designed as a tool for analysis of distributed knowledge inside companies and organizations, incorporates test-based dynamics and several other useful features that simplify management of accounts, presentation of questions and collection and analysis of answers. As it was established, the proposed methodology has been particularized in a proof of concept for the Manufacturing Processes subject at EPIG engineering degrees. In order to create the structure of campaigns that the ZAPIENS platform uses, four blocks have been defined: Standardization and Tolerances; Forming and Shaping Processes; Metal-Casting Processes and Machining Processes. Thirty test questions have been included in each block, so that each question is given three possible answers (only one of them true). Consequently, one hundred and twenty test questions shall be answered by each student during the whole semester. An example of an actual question can be seen on figure 2.
ZAPIENS allows decisions on how the test shall be presented to the students. In this work, each campaign starts after the theoretical base for its corresponding block of knowledge has been completed in classroom during the theory sessions, and each student should answer three questions every day of a ten-day period. The platform also provides real-time feedback on the achievements of the students with respect to those of their mates, since a rank of achievements is constantly updated. This rank serves to the purpose of motivating the students to improve their performances. Similarly, the platform provides information about the behaviour of students. The administrator has access to an enriched set of data that includes information on when the students are connected, how long they took to answer a particular question or the percentages of failure for each answer.

2.4. Extension of research
A total of 297 students were enrolled in the subject during the academic year 2017-2018. To avoid possible crossed-influences, it was decided that the analysis would be limited to those who take the subject for the first time. Consequently, the number of students is reduced to 110. Then, those students who had answered less than 100 questions are also supressed from the study, in order to avoid biases between campaigns due to asymmetric participation. Finally, the initial figure is reduced to 92 students. On the other side, a total of 334 students were enrolled in the subject during the academic year 2018-2019. Under the same considerations as before, 68 students that take the subject for the first time at least 100 of the total of 120 questions proposed. In both cases, regarding the final evaluation, it was established that a successful student would pass the course during the academic year, no matter if this goal was accomplished in January or June calls.

3. Results and Discussion
First of all, it must be clarified that, although the two-parameter IRT model had been selected, based on previous experiences, other well-known commonly employed IRT models, like the one-parameter and three parameters [12] were also tested. Nevertheless, they provided worst results according to Akaike Information Criterion and Bayesian Information Criterion [13]. Therefore, these alternative models were rejected for this case study and the following results were conducted using exclusively the two-parameter model.

![Figure 3. Boxplots of the ability level obtained through IRT against the final achievement of the student obtained from each Block–Course 2017-18.](image)

### 3.1. Statistical significance of ability within the 2017-18 course
Once the campaigns have been completed and the 2017-18 course has already finished, results of ability processed through IRT and final result (pass/fail) can be used to check if there is a significant relationship between both. Firstly, each campaign or BLOCK has been processed independently. Then, for each student and each campaign a unique value of ability is obtained. Figure 3 contains boxplots of
the values obtained for the ability parameter in each campaign against the achievement of the student at the end the course (pass: yes/no); the value for the median parameter is also given in each boxplot. Although all campaigns seem to provide similar results, with students who had fail to pass the course showing comparatively lower values of ability with respect to their mates who actually had succeed, the level of significance of this relationship between blocks does not seem to be equally clear. If a statistical analysis of this correlation is carried out, results indicate that the ability calculated from BLOCK1 and BLOCK4 have a significant relationship with the final result, whereas BLOCK2 and BLOCK3 are not. Table 1 contains the results of statistical two-sample t-test.

Table 1. Results of statistical two-sample t-test for the ability parameter calculated for each block of knowledge.

|       | Passed | No | Yes |
|-------|--------|----|-----|
| N     | 48     | 44 |     |
| Mean  |       |    |     |
| St. Dev. |    |    |     |
| BLOCK1 | 0.409 | 0.774 | 0.261 | 0.809 | 0.000 |
| BLOCK2 | 0.161 | 0.893 | 0.119 | 0.852 | 0.128 |
| BLOCK3 | 0.111 | 0.957 | 0.074 | 0.774 | 0.308 |
| BLOCK4 | 0.366 | 0.925 | 0.116 | 0.822 | 0.010 |

Statistical significance has also been analyzed by comparing the final results with an ability parameter calculated for all the campaigns taken as one single block of knowledge. Figure 4 shows the corresponding boxplot with the median values for each group. A two-sample t-test has been also performed for the complete course campaign, and it has revealed that the ability values obtained in the analysis of the four blocks for those students that failed the exam were significantly lowers than the values of those who passed the exam. It means that those students that obtained higher ability levels in the test questions are more prone to pass the subject. Nevertheless, the p-value calculated in this test is higher than those calculated for BLOCK1 and BLOCK4. This means that it is not necessary to complete the sequence of campaigns to have an adequate estimation of the possibilities that each individual student has to finally pass the course. In fact, since BLOCK1 campaign is driven at the beginning of the course, it seems feasible to adopt a strategy for early identification of high risk-of-failure students based on the ability level calculated from this particular campaign.

3.2. Discriminating ability values calculation (2017-18 course)

BLOCK 1 has been selected as an early indicator of success, and the calculation of $\theta_{HRF}$ and $\theta_{HPP}$ have been done based on data extracted. Based on the number of students in the experiment during the 2017-18 course, LAG and HAG should contain 18 students (19.56%). Once the size of each group has been fixed, the $\theta_{HRF}$ would be that corresponding to the 18th student, given an increasing ability order.
Accordingly, $\theta_{HRF}$ will be fixed at a -0.696 value, and students scoring this value or lower would be assigned to the LAG. Similarly, the $\theta_{HPP}$ would be that corresponding to the 18th student, given a decreasing ability order. Therefore, $\theta_{HPP}$ will be fixed at a +0.761 value, and students scoring this value or higher would be assigned to the HAG. Figure 5 contains two graphics where the percentage of accurate assignment of students to LAG and HAG is presented against the level of ability used as cut point for courses 2017-18 and 2018-19.

It can be noticed in Figure 5, for the No Pass, selecting very low values for $\theta_{HRF}$ would provide a 100% accurate assignment. As the selected $\theta_{HRF}$ increases, incorrect assignments and accurate ones alternate, which causes the percentage to fluctuate. At the selected number of students (18), there will be 14 students accurately clustered (failing to pass the subject) whereas the other 4 have finally passed the course. It seems reasonable that the accuracy of clustering strategy for LAG reaches such level in order to future developments of reinforcement strategies oriented to the students in this group. Similarly, for the Pass, also shows that very high values for $\theta_{HPP}$ would provide a 100% accurate assignment. As the value of $\theta_{HPP}$ decreases, incorrect assignments appear and the percentage of accuracy drops.

Considering the desired number of students (18) in the group, there will be 15 students accurately clustered whereas the other 3 have failed to pass the course.

3.3. Accuracy evaluation based on results from the 2018-19 course

During the following academic year (2018-19) test campaigns were driven again within the same conditions. When the course was finally done, final results of the students were compared with the ability parameter calculated from the first (BLOCK1) campaign. Results are also given in Figure 5.

Considering the values of $\theta_{HRF}$ and $\theta_{HPP}$ calculated during the previous course, 16 students would be clustered in the LAG for the 2018-19 course ($\theta \leq -0.696$) and, given the final results, the level of accuracy obtained by supposing that they would fail the exam was 75%. Similarly, 12 students would be assigned to the HAG ($\theta \geq 0.761$) and the accuracy of supposing that they would pass the exam was 83.3%.

Consequently, using the limits for the ability parameter that had been calculated during the precedent academic course would lead to similar values of accuracy in the assignments to LAG and HAG. Although the number of students within each group is unbalanced, the strategy shows its capability of correctly clustering three of every four students in the LAG and four of every five in the HAG. These results could be the key to the introduction of specially-designed strategies regarding reinforcement and adapted learning as depending on which group the student has been assigned to.

4. Conclusions

In present work, a methodology for early identification of those students that are more likely to fail a subject has been proposed. Additionally, a proof of concept that considers the particular case of a manufacturing engineering subject during the courses 2017-18 and 2018-19 has been carried out.
Batteries of test grouped according to blocks of knowledge that had been previously defined for the current subject have been presented to the students using the ZAPIENS on-line knowledge management platform. Results from this test blocks had latter been exported and processed in R by means of the IRT. Considering the 2017-18 course, it has been demonstrated that results of the ability parameter calculated for certain blocks of knowledge could provide a statistically significant relationship with the final result (pass/fail) of the course. In the particular case of the proof of concept, BLOCK1, a transverse content related to standardization and tolerances, was considered. Therefore, ability values calculated using IRT upon data collected from this block of knowledge has been used to classify students into three groups: HAG (highest abilities), LAG (lowest values of ability) and MAG (the rest).

The capability of this early indicator has been finally tested upon the students in 2018-2019 course, and the analysis indicate that this procedure provides good results as an early clustering method. Thus, when no more than one fourth of the semester has passed, faculty has the possibility of identifying a group of students in risk of failing the subject, when there is still enough time left for applying reinforcement measures - specially focused on the LAG - that could help those students to keep connection with the subject and, in the end, pass the subject. Additionally, those students in the HAG could be also offered specially-designed contents. This methodology could be easily adopted, no matter the academic level or the characteristics of the course. In future work, the intention is to propose and evaluate possible strategies, with the final objective of improving efficiency figures: ER, SR and, eventually, DR.

Acknowledgements
The authors wish to thank the ZAPIENS team and the Teaching Innovation Committee of the University of Oviedo for their support.

References
[1] European Commission 2015. Dropout and completion in higher education in Europe. Available: https://doi.org/10.2766/826962
[2] Armenteros J H, Pérez J A 2018 La Universidad Española en Cifras 2016/17. (Conferencia de Rectores de Universidades Españolas)
[3] Eamon M K 2005 Social-demographic, school, neighborhood, and parenting influences on the academic achievement of Latino young adolescents J. Youth Adolescence 34 (2) pp 163–174
[4] Guio J, Choi A, Escardíbul J O 2018 Labor markets, academic performance and school dropout risk: evidence for Spain, Int. J. Manpower 39 (2) pp 301–318
[5] Yang S. J. H et al. 2018 Predicting students' academic performance using multiple linear regression and principal component analysis, J. of Inform. Proces. 26 pp 170–176
[6] Meier Y et al. 2015 Predicting grades, IEEE T. Signal Proces. 64 (4) pp 959–972
[7] Mi C et al. 2018 A dynamic early warning method of student study failure risk based on fuzzy synthetic evaluation, Int. J. Perform. Eng. 14 (4) pp 639–646
[8] Bortolotti S L V et al. 2013 Relevance and advantages of using the item response theory, Qual. Quant. 47 (4) pp 2341–2360
[9] Gulliksen H 1961 Measurement of learning and mental abilities Psychometrika 26 (1) pp 93-107
[10] Rizopoulos D 2006 ltm: an R package for latent variable modeling and item response theory analyses J. Stat. Softw. 17 (5) pp 1–25
[11] Ayala R J 2009 Methodology in the social sciences. The theory and practice of item response theory (New York, NY, USA Guilford Press)
[12] Lord F M 2012 Applications of item response theory to practical testing problems (New York, NY, USA: Routledge)
[13] Akaike H 1974 A new look at the statistical model identification IEEE T. Automat. Contr. 19 (6) pp 716–723