ADAPTATIVE TESTS AS AN EVALUATION METHOD IN THE STEM CONTEXT: AN EXPERIENCE IN THE ENERGY TECHNOLOGY DOMAIN

J.J. Serrano-Aguilera, J. Prieto, J.P. Jiménez-Navarro, C. Martin, A. Tocino

University of Malaga (SPAIN)

Abstract

The work developed in this study focuses on improving evaluation methods implemented in the Energy Technology course of the Master's Degree in Industrial Engineering at the University of Málaga (Spain). For this purpose, adaptive assessment techniques, which are common in other teaching fields, such as language teaching, were developed and analysed. The potential benefits of this new tool in the field of STEM education mainly rely on the improved ability of adaptative tests to measure the level of acquisition of skills and knowledge by the student in less time and with less consumption of resources. This advance facilitates a continuous and detailed monitoring during evaluation tasks in the learning process. This work, organized in two stages, has first analysed the role played by several factors such as the perception of difficulty by the student and by the lecturers, the response time and the success rate of the questions. Subsequently, and based on the previous information, adaptive tests were designed based on decision trees that consider two essential criteria: the difficulty level of the question and the thematic area that each question belongs to within the syllabus. Results indicate that adaptive tests show a significant reduction in the time invested in the tests without substantially altering the evaluation results and grades obtained.

Keywords: Adaptative tests in Engineering, STEM Education, Decision tree

1 INTRODUCTION

The assessment process is one of the fundamental processes during student learning. It serves two important purposes: the feedback that students receive can help the process [1], and the verification that certifies whether the student has acquired the expected competences. It has been observed that assessment employing test-based examinations allows a wide range of concepts to be covered, reduces marking times, as well as conveys a greater perception of objectivity during the revision process, among others [2]. Computer adaptive tests can automatically adapt the exam by selecting the next question according to previous answers [3]. This approach introduces adaptability in educational systems, where learners can adapt the course material by themselves [4]. Based on a learner model, the main goal of this approach is to precisely estimate the student's abilities while reducing the total number of questions required [5]. Adaptive assessment can positively impact students by reducing test anxiety, increasing engagement, and providing more efficient and personalized evaluations. Recent approaches consider the use of Artificial Intelligence (AI) for personalized learning experiences [6], whereas traditional approaches mainly consider Item Response Theory (IRT) [7]. However, in the context of STEM (Science, Technology, Engineering and Mathematics) this type of assessment has not had a wide impact, unlike other branches of science, such as Economics, where it is estimated that between 45 and 67% of the assessment processes are carried out in this way [8]. One of the main reasons for this is the difficulty in assessing certain concepts taught in this type of subject, such as the creation of an algorithm in the field of computer science, which often requires complex problem-solving exams or even the development of laboratory sessions. This project explores the application of an evaluation methodology with extensive experience in the field of language learning [9] to one subject in the field of STEM education.

The main objective of this research is to develop an experience towards the improvement of assessment processes. This project aims to deepen essential concepts known by students during the assessment process, while pointing out those aspects for which students lack sufficient preparation. It is therefore a win-win project in the educational context: 1) it will allow teachers a more personalised assessment; 2) students will be able to combine their efforts in a self-guided way in the subject with better preparation; and 3) it can reduce assessment times and stress for students, and therefore for teachers; 4) in the

context of multiple-choice tests, the risk of false positives is reduced, which provides a more realistic grade.

The rest of the paper is organized as follows. Section 2 presents the methodology used for the assessment of this work. In Section 3, the main results of this work are discussed. Finally, conclusions are outlined in Section 4.

2 METHODOLOGY

This study implemented a non-randomized adaptive testing approach to evaluate the acquisition of competencies by students in the *Energy Technology* course, part of the Master of Industrial Engineering at the University of Malaga. The methodology aimed to assess how adaptive assessments influence the learning and evaluation processes, as well as to determine whether students perceive this method as valuable for future academic use. The study was conducted over two academic years (2022–2023 and 2023–2024), during which adaptive tests were progressively designed, implemented, analyzed and compared with traditional non-adaptative tests. These tests were conceived as tools not only for grading purposes but also for pedagogical diagnosis, enabling the identification of conceptual gaps and areas where students required reinforcement.

The main principle underlying the adaptive tests was the use of a decision-tree model to personalize the assessment path based on student responses. The tests were structured as sets of multiple-choice questions, hierarchically organized according to their conceptual depth and thematic relevance. Each thematic block in a course was associated with a dedicated sub-tree. If a student answered a question correctly, the system would guide them toward more complex questions within the same conceptual framework. Conversely, incorrect answers would result in either a regression to more elementary items or a shift to another thematic area. This dynamic structure was designed to reduce the number of redundant questions and to optimize the assessment process by focusing on the most informative items for each student's competency level.

Two main criteria informed the construction of the question bank: (i) the difficulty level of each question, as defined by the course lecturer based on pedagogical relevance and cognitive complexity; and (ii) the association of each item with a specific lesson or module within the course syllabus. This dual classification enabled a detailed adaptation mechanism during the testing process, ensuring both thematic coherence and cognitive progression.

The adaptive tests were developed using **SIETTE**, a web-based system created at the University of Malaga (https://www.siette.org/siette/). SIETTE facilitates the creation and maintenance of extensive question repositories and supports a variety of assessment strategies, including adaptive testing algorithms. The platform allows for controlled access to evaluations, customizable selection criteria, and real-time monitoring of test completion. It is fully integrated with the University's Virtual Campus, which enabled a seamless deployment of the assessments in real course environments. Furthermore, SIETTE provided robust data logging capabilities that were crucial for tracking student performance and analyzing response patterns.

This study was structured in two stages which were conducted throughout two consecutive academic years:

First stage. Academic year (2022-2023): Development of tools and structures of adaptive tests. The number of difficulty levels (an essential criterion for developing adaptive tests) and their correspondence with the students' perception of difficulty have been analyzed. It was also required to focus building the initial question bank specifically for the *Energy Technology* course. Lecturers defined the hierarchical structure of basic and advanced concepts and conceived the general layout of decision trees to be implemented on the second stage. At this stage, student feedback was actively collected through preliminary implementations to adjust question wording, difficulty calibration, and navigation flow of SIETTE platform.

Second stage. Academic year (2023-2024): By the end of the first stage, a complete framework of decision trees and question taxonomies had been finalized, enabling a more extensive deployment during the 2023–2024 academic year, still within the context of the *Energy Technology* course. Students

completed adaptive assessments based on a curated set of 80 multiple-choice questions distributed across four thematic blocks (topics). Each student was presented with a customized set of questions depending on their individual responses, and the system recorded detailed metrics including answer accuracy, response time, and self-reported difficulty (on a discrete scale from 1 to 4). This results we also compared with those obtained after completing a “linear test”, which consisted of going through the whole 80 multiple-choice tests set but in a sequential order (non-adaptative approach).

To enable a rigorous comparison between adaptive and non-adaptive formats, a follow-up assessment was administered in which students completed the entire question set. This allowed researchers to analyze differences in test length, grading accuracy, and conceptual coverage between the two approaches. The adaptive model was structured on the assumption that a correct response to a difficult question indicates mastery of simpler underlying concepts, thereby justifying the omission of redundant items. This logic aimed to reduce both cognitive load and completion time for proficient students.

3 RESULTS

In this section, we present the quantitative results of the adaptive assessment under real evaluation conditions. In a first stage, test organised in June 2023, students were asked to answer a conventional multiple choice test that was used for two main purposes: first, to evaluate the robustness of the in-house developed tool (SIETTE) under real test conditions and second, to understand the perception of difficulty in the questions answered by students. To this end, students were asked to rank questions in four levels of difficulties during the test (see Figure 1). These replies were compared with the difficulty previously assigned by lecturers. In this test, we also recorded the response time to check if students’ difficulty perception was correlated to the time that they spent to provide a reply. The test was composed of 60 questions and no adaptative approach was implemented. It means that all questions where answered in a predifined squential order (named as *non-adaptative* or *linear test*). A total of 28 students completed the test, including answers to all questions and difficulty perception.

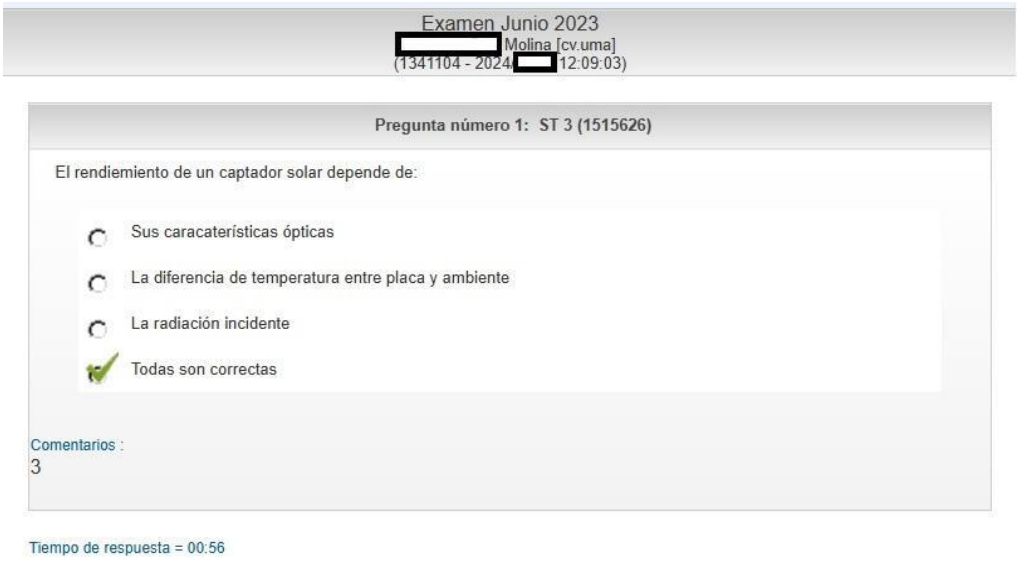

Figure 1. Anonymised snapshot of SIETTE graphic interface used in the test (Spanish text). Three variables of interest have been recorded: (i) response time, (ii) a small text box, where the student must type the level of difficulty he/she considers for the question on a scale of integers (1-4): 4 indicates maximum complexity, whereas 1 corresponds to the least difficulty, (iii) the student's answer to the question

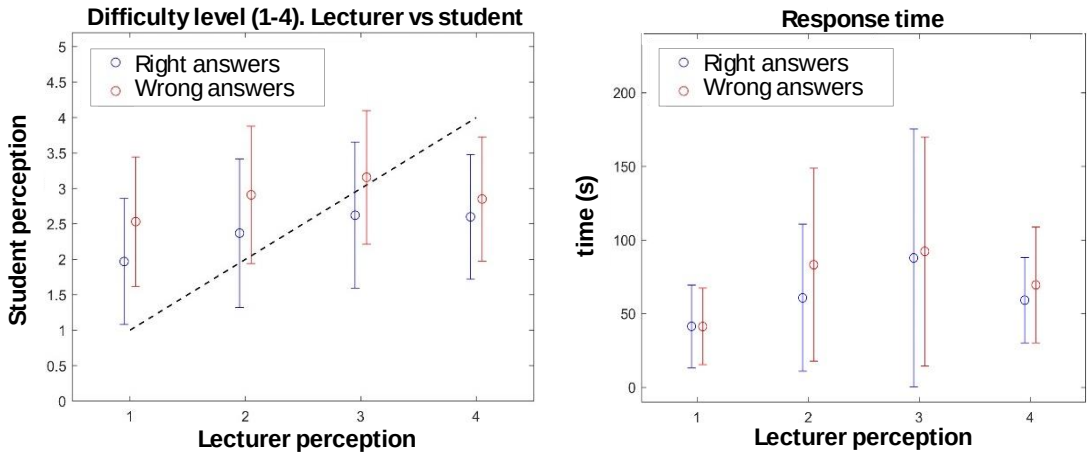

Figure 2. Correlation between the level of difficulty established by the professor and the student (left) and correlation between the level of difficulty established by the lecturer and the response time (right).(NB: Results grouped by right and wrong answers)

The compilation of all the previous results were suitable to validate the question classification criteria, which is a key factor in the development of an adaptive test. The questions were classified according to their level of difficulty (scale 1-4) and whether the answers provided by the student were correct or incorrect (right vs wrong answers). As depicted in Figure 2-left, even though students perceive higher levels of difficulty compared to the lecturer, a direct correlation is observed for the first three levels of difficulty (1-3). Similar trend is reported when comparing with the response time (Figure 2-right). A higher level of difficulty set by the lecturer leads to longer response times. As expected, more complex questions result in lower right answer rates as presented in Figure 3.

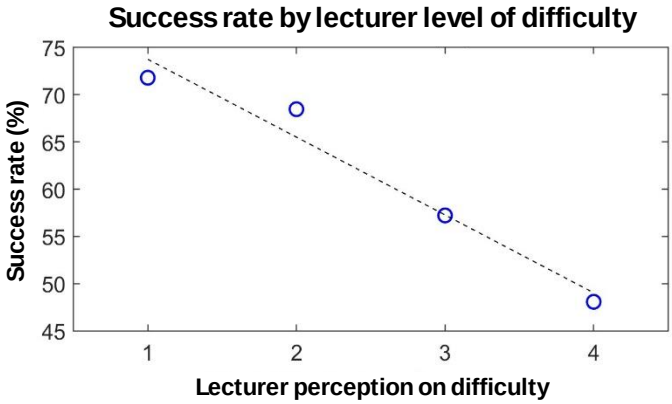

Figure 3. Average success rate of the students to the level of difficulty of the question (scale 1-4).

Particularly interesting are the results for questiones tagged as high difficulty (level 4). As displayed in Figure 2, correlation is lost for this set of difficult questions. Explaination for this may lay with the student distorted perception particularly for those questions with a high degree of difficulty. This may be because students pick a random answer (wrong answers do not affect the final mark) to save time and focus on questions with less difficulty. This particular phenomenon has had an impact on both: the response time and the perception of difficulty by students (Figure 2). From this first stage, the main conclusion that can be drawn is that the level of difficulty assigned by lecturers, although subjective, is an adequate criterion based on its correlations with the three variables considered: student perceptions, lecturer perceptions, and response time. It seems clear that in the development of an adaptive test, it is always advisable to conduct a previous study based on the statistical analysis of several factors, as well as their correlations. It is not ruled out that there may be new factors that improve this process, but so far, the three factors

recorded have shown their ability to give some insight about the suitability of the classification and criteria followed.

Te second stage, implemented during the academic year 2023-2024, included the adaptative approach that was built upon the conclusions from first stage. As explained in the previous section, results are compared between a linear test and an adaptative tests that students are required to sit. In both tests, each thematic area (topic within the syllabus) is marked independenlty and following a sequential order.

The structure of the adaptive test is based on a decision tree, as shown in Figure 4. Each thematic block has its own tree and replicates the same structure. First, the student is asked the simplest questions (level 1). It must be passed before moving on to level 2 (within the same thematic block). For this, if he/she answers two questions correctly (e.g. 1A and 1C), he/she goes directly to the next level. Otherwise, the number of questions increases in order to pass to the next level or even fail the entire thematic block "END". It should be noted that there are 4 levels of difficulty, and the grade obtained includes all answers given during the path followed through the entire decision tree.

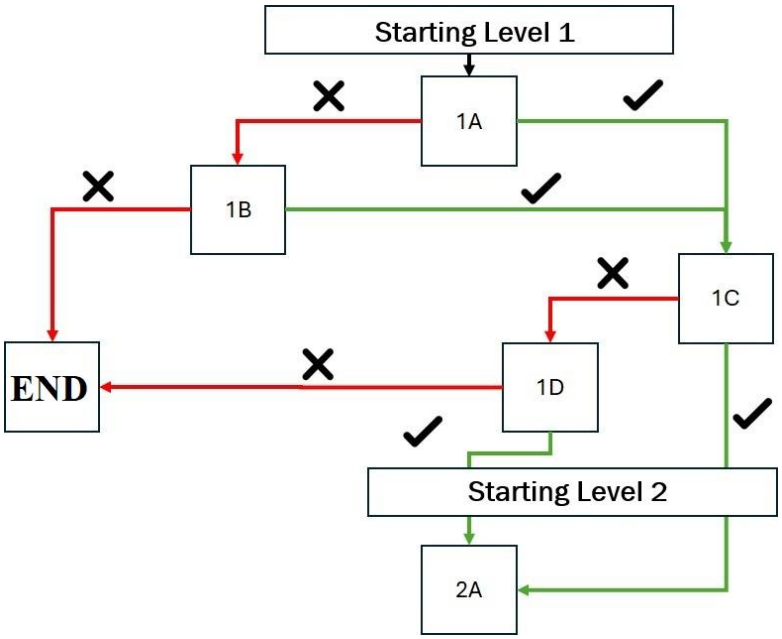

Figure 4. Decision tree layout for level 1 and a specific thematic block. The complete scheme is a repetition of this scheme for each level and each thematic block. The question code "1A" indicates question 'A' of difficulty level "1". Although the database has five questions for each level 1A-1E, only 4 are used in the adaptive test 1A-1D. The red lines (symbol X) are those paths that are followed after a wrong answer. The green ones (tick) in the case of a correct answer.

Results shown in Figure 5 depicts correlation between results of the linear and adaptative tests for the students that sat the assessment and completed both tests. Data points that fall close to the unitary slope line are those that show a better correlation between the linear and adaptative test. This is, a similar mark was obteined in both tests, which supports the use of adaptative tests and an alternative and time efficient way of evaluating students. The weights given to each question according to its level of difficulty represent a set of adjustment parameters that can be optimized to improve the predictive potential of adaptive tests. However, those weights do not seem to have a significant impact. If we compare results in Figure 5-left (each weighting parameter has a value of 0.25) to those in Figure 5-right, where weights were set to favour more difficult questions (w_1 = 0.2 and w_4 = 0.3) similar distributions are observed. The second option provides a slightly better adjustment since the MSE falls by 1.7% from 114.41 to 112.42. In both cases, the set of points is distributed along the unitary slope line, which indicated that the adaptative tests, despite some level of error, clearly provides a fit trend.

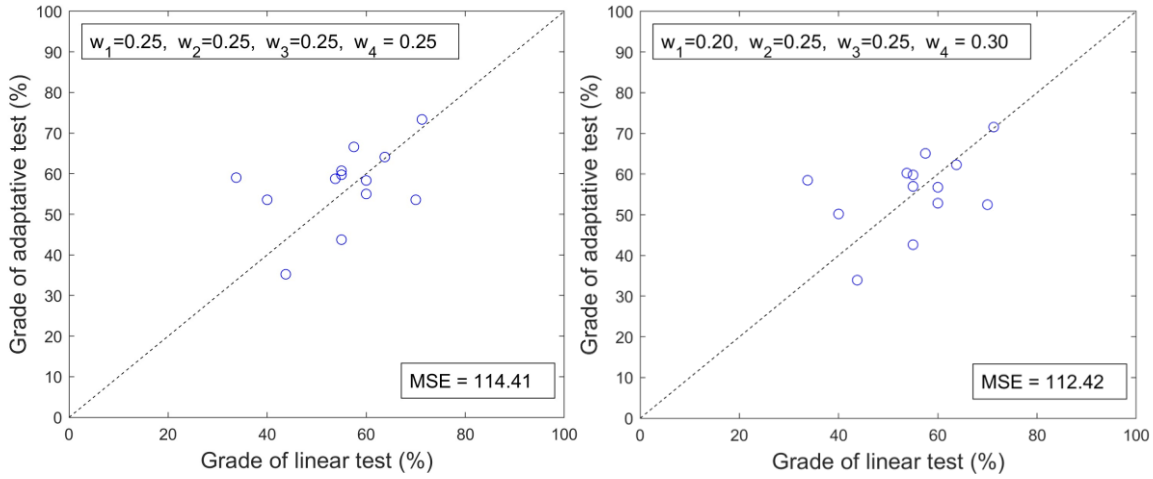

Figure 5. Relationship between the score obtained by the linear test and the adaptive test where all questions have been given the same weight (left) and between the score obtained by the linear test and the adaptive test where different weights have been given to the questions according to their 1-4 level of difficulty (right). MSE stands for the Mean Squared Error respect to the unitary slope.

In order to provide a quantitative measure of the time-saving potential of the adaptative evaluation, Figure 6 presents the number of questions completed in each adaptative test. The number of questions in the adaptative test is considerably lower than those in the linear test (80 questions). This shows the potential of adopting adaptative tests to save time in the marking process. However, as expected, the major the mark in the examen, the larger the number of questions to be answered. This is primarily due to the fact that low-grade tests may have skipped part of the high-difficulty level blocks.

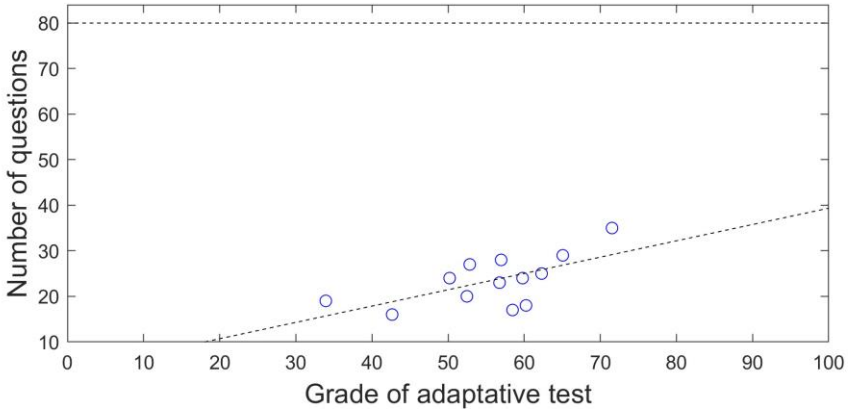

Figure 6. Distribution of the number of questions answered according to the score obtained in the adaptive test. The linear test consists of 80 questions (dashed horizontal line).

4 CONCLUSIONS

The results obtained from this project demonstrate the potential of assessing STEM syllabus using adaptive techniques as the one presented in this work. In particular, this approach can replace traditional tests optimising the marking time and providing effective understanding of student´s knowledge. Yet, the implementation of such approach requires adequate tools, such as the SIETTE tool, developed at the University of Málaga, that has helped us evaluate simple adaptive techniques, and evaluate the adequacy of such approaches.

Our analysis suggests that, to develop adaptive tests, it is advisable to plan a structured classification of questions grouping them into the thematic areas covered in the syllabus and level of difficulty. To ensure an adeaquate clusterisation of questions, it is important to assess the correlation between lecturers' and the students' perception of the difficulty of the questions in the test. Response times are also an indicative variable to be considered. Except for the most complex questions, for which student perception is distorted, there is a strong correlation with the rest of the variables. This is, in the most difficult questions (level 4), students do not necessarily spend more time answering them.

After implementing the adaptive tests in the second stage of the study, it can be reported that the reduction in the number of questions and the overall time of the test are significant, without substantially affecting the grade in the test with respect to the linear tests. This concordance was better for students with high scores. Last, we analysed the effect of setting variable weights on the final grade for each question according to their level of complexity. However, we cannot conclude that this sort of adjustment leads to any relevant improvement in the concordance mentioned above.

ACKNOWLEDGEMENTS

This study has been developed in the framwork of the INNOVA22 program of the University of Malaga, through which the authors have implemented the Teaching Innovation Project with reference number **PIE22-085**. Authors would also like to acknowledge the funding provided by Universidad de Málaga (UMA) as well as the support provided by the creator of the SIETTE tool, *Dr. Ricardo Conejo*, professor of the Department of Languages and Computer Science at Universidad de Málaga. In addition to solving our needs, he has made modifications to the SIETTE tool to achieve our objectives. We would like to thank him warmly for his help and work.

REFERENCES

[1] C. Evans, "Making Sense of Assessment Feedback in Higher Education", Review of Educational Research 83, 1, pp. 70–120, 2013.

[2] M.G. Simkin & W.L. Kuechler, "Multiple-choice tests and student understanding: What is the connection?", Decision Sciences Journal of Innovative Education, 3(1), pp. 73-98, 2005.

[3] VIE, Jill-Jênn, et al. A review of recent advances in adaptive assessment. Learning analytics: Fundaments, applications, and trends: A view of the current state of the art to enhance e-learning, 2017, p. 113-142.

[4] Van der Linden, Wim J., and Cees AW Glas, eds. Elements of adaptive testing. Vol. 10. New York: Springer, 2010.

[5] Lan, A. S., Waters, A. E., Studer, C., & Baraniuk, R. G. (2014). Sparse factor analysis for learning and content analytics. The Journal of Machine Learning Research, 15(1), 1959-2008.

[6] Gligorea, Ilie, et al. Adaptive learning using artificial intelligence in e-learning: A literature review. Education Sciences, 2023, vol. 13, no 12, p. 1216.

[7] Van der Linden, Wim J., et al. (ed.). Computerized adaptive testing: Theory and practice. Dordrecht: Kluwer Academic, 2000.

[8] W.E. Becker & M. Watts, "Teaching methods in US undergraduate economics courses", The Journal of Economic Education, 32(3), pp. 269-279, 2001.

[9] M.M. Hicks, "The TOEFL computerized placement Test: Adaptative Conventional Measurement", ETS Research Report Series, pp. i-29, 1989.