Tendências em Matemática Aplicada e Computacional, **20**, N. 2 (2019), 381-401
© 2019 Sociedade Brasileira de Matemática Aplicada e Computacional
www.scielo.br/tema
doi: 10.5540/tema.2019.020.02.0381

Academic English Proficiency Assessment Using a Computerized Adaptive Test

M. CÚRI[1*] and V. SILVA[12]

Received on January 21, 2017 / Accepted on March 13, 2019

ABSTRACT. This paper describes the steps of transforming a paper-and-pencil English proficiency test into an computerized adaptive test (TAI-PI) based on an Item Response Theory (IRT) model. The exam is composed of multiple choice items administered according to the Admissible Probability Measurement Procedure, adopted by the graduate program at the Institute of Mathematics and Computer Sciences at the University of São Paulo (ICMC-USP). Despite the fact that the program accepts various internationally recognized tests that attest non-native speakers English proficiency, such as the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS) and the Cambridge English: Proficiency (CPE), for instance, its requirement is incompatible with the way the Brazilian public university operates due to the cost, which ranges from US$ 200.00 to US$ 300.00 per exam. The TAI-PI software (Computerized Adaptive Test for English Proficiency), which was developed in Java language and SQLite, started to be used to assess the English proficiency of students on the program from October, 2013. The statistical methodology used was defined considering the history and aims of the test and adopted Samejima's Graded Response Model, the Kullback-Leibler information criterion for item selection, the a posteriori estimation method for latent trait and the Shadow Test approach to impose restrictions (content and test length) on the test composition of each individual. A description of the test design, the statistical methods used, and the results of a real application of TAI-PI for graduate students are presented in this paper, as well as the validation studies of the new methodology for pass or fail classification, showing the good quality of the new evaluation system and examination of improvement using the IRT and CAT methods.

Keywords: computerized adaptive testing, item response theory, shadow test.

1 INTRODUCTION

Having a good command of English is fundamental for graduate students from various fields of science, so that they will be able to understand the course content properly, as well as develop and disseminate research carried out. Taking this into account, many graduate schools in

*Corresponding author: Mariana Cúri – E-mail: mcuri@icmc.usp.br – https://orcid.org/0000-0002-7651-1064

[1] Universidade de São Paulo, São, Paulo, SP, Brazil. E-mail: mcuri@icmc.usp.br, varufino@gmail.com

[2] Universidade Federal de São Carlos, São Carlos, SP, Brazil.

Brazil require proof of English proficiency from their students, either to enroll on the program, or throughout the course.

There are various internationally recognized tests, which are generally accepted attesting non-natives' English proficiency. Among the most traditional, the following can be cited: Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS) and the Certificate of Proficiency in English (CPE). However, one of the biggest drawback of these tests is the cost, which ranges from approximately US$ 200.00 to US$ 300.00 per exam. This is one of the main reasons why the graduate program in Computer Science and Computational mathematics (CCMC in Portuguese) and Statistics (PIPGEs in Portuguese) at ICMC-USP (the latter having a joint program with the Federal University of São Carlos (UFSCar)) offer a Proficiency Test in English called EPI as a free alternative for their students.

Until 2013, the EPI was a paper-and-pencil test and the aim of this paper is to make it a Computerized Adaptive Test (CAT), in which the items are selected gradually for the individual, according to their proficiency. The value of the proficiency of each individual is updated after each answer, based on Item Response Theory (IRT) model [2]. Therefore, very easy or very difficult items for a given individual are not even presented to the candidate, reducing the test length and performance time [26], making it more efficient, objective and producing an immediate result [34]. Figure 1 shows the item selection and administration procedures of the test and update of proficiency estimates, iteratively, until some stopping criterion is met.

Besides the advantages mentioned above, CAT in combination with IRT make it possible to calculate comparable proficiencies between individuals who answered different sets of items, and at different times [14, 32]. This greatly facilitates evaluating constructs on a large-scale resulting in its use in important examinations, such as the Graduate Record Examination (GRE) [6, 11], developed by the Educational Testing Service (ETS) in 1996; the TOEFL [10, 12, 33], also developed by ETS and the Armed Services Vocational Aptitude Battery Test [23, 24], developed by the United States Department of Defense to select potential recruits for military service.

To justify the choice of statistical methodology adopted in this study, in what follows in this section, we present the format in which the EPI was applied to the CCMC program between 2002 and 2013, providing the basis for building the Computerized Adaptive Test of English Proficiency (TAI-PI).

In section 2 is presented the characteristics and structure of the test applied to students. In section 3, the computerized adaptive testing methodology implemented in TAI-PI is described. Section 4 presents some simulation studies to evaluate the efficiency in latent trait estimation comparing different test lengths and different prior distributions assumed in Expected a posteriori (EAP) bayesian estimation method. The results of the real application of the test with the created system (TAI-PI) are described in section 5 and a classification study is conducted based on these results in order to establish a cut-off point in the latent trait scale. Finally, conclusions are drawn in section 6.

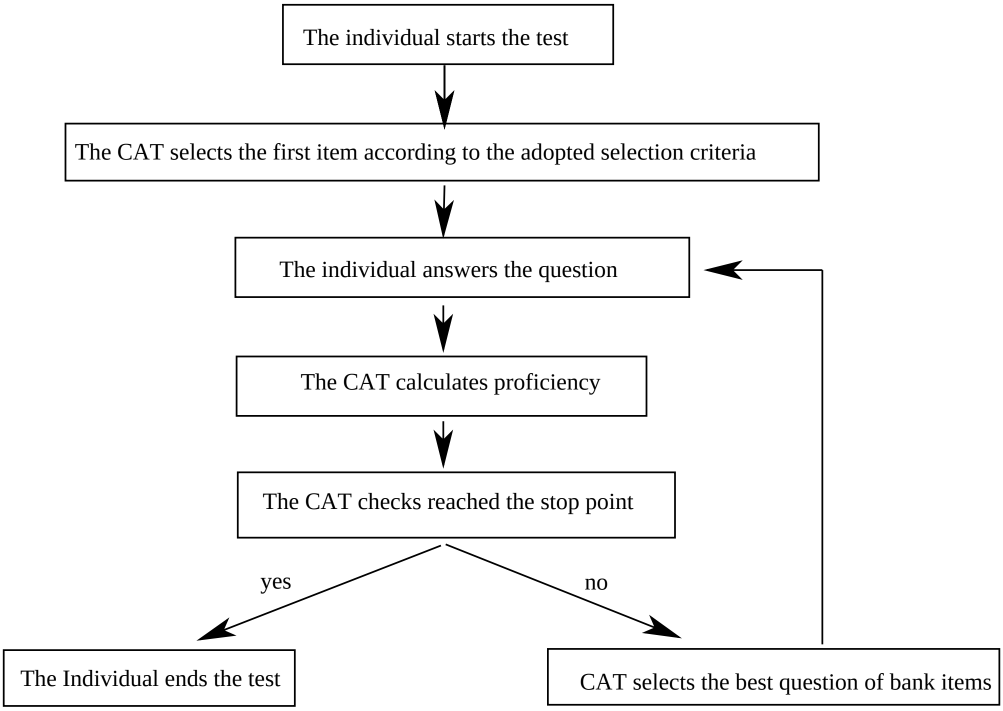
The individual starts the test
The CAT selects the first item according to the adopted selection criteria
The individual answers the question
The CAT calculates proficiency
The CAT checks reached the stop point
yes
no
The Individual ends the test
CAT selects the best question of bank items

Figure 1: Diagram of a computerized adaptive test.

2 EPI

Since 2002, the EPI has consisted of a multiple-choice item test, offered every six months in paper-and-pencil format. To create the test, 25 to 30 item were selected from a database of 167 items in total, divides into three modules as follows:

1. In Module 1, abstracts from scientific journals in the fields of Computing, Applied Mathematics or Statistics and questions about the components of **scientific structure** are presented.

2. In Module 2, there is part of an introduction of a scientific paper from the same areas, as well as question (in Portuguese) about **reading comprehension** and the relationship between ideas in it.

3. In Module 3, the questions are related to **grammar conventions of language**, such as conjunctions, verb tenses, relative clauses and articles.

The items form the database were prepared by two university professors (one who is and English Language professor at the Federal Institute of Rio Grande do Sul, and the other a professor of Computer Science at ICMC-USP). The development was based on Bloom's Taxonomy [5] aiming to evaluate competence in reading and understanding scientific papers in the fields of Computing, Applied Mathematics and Statistics.

In order to minimize the possibility of "guessing" the correct answer and not assuming that student's knowledge about a particular item is only binary (correct or incorrect) the classification of the student's response follows the Admissible Probability Measurement Procedure (APM), [13, 25]. Figure 2 illustrates the possibilities of an examinee's answer to an item from the EPI. There are three alternatives for each multiple-choice item (A, B and C), represented at the vertices of an equilateral triangle, and only one if them is correct. However, the student has 13 possibilities of answering, which express the extent of his/her certainty about which of the three statements is correct. This can be selected in the following way:

- If the examinee is sure which is the correct answer, he/she should use options A, B or C.
- If the examinee does not know the correct answer and is totally uncertain, he/she should use option M.
- If the examinee is in doubt between options A and B (equally likely), he/she chooses the option E. Analogously, he/she can choose options H (when the doubt is between options B and C), or K (when the doubt is between options A and C)
- If the examinee is in doubt between two options, but one of them seems to be more correct than the other, he/she chooses the point that is closest to the preferable alternative (D or F, if the doubt is between alternatives A and B, for instance). Similar interpretation can be applied for points G and I (between alternatives B and C) and J and L (between alternatives A and C).

In figure 2 there is the student response classification according to his/her choice: fully informed, informed, partially informed, misinformed, totally misinformed and uninformed [1].

The student passes if the percentage of answers classified as "fully informe" is greater than or equal to 50% and if the percentage of "totally misinformed" answers is less than or equal to 25% or 90% or more of the answers are in the "fully informed", "informed" and "partially informed" classes and 10% or less of the answers in the "totally misinformed" class. For students who failed according to this criterion, but almost passed if it was not for the answers of one or two items (maximum), an alternative criterion is available considering only Module 1 "Scientific Text Structure" .

3 TAI-PI METHODOLOGY

The TAI-PI was developed in Java 1.7.0 and SQLite, version 3.7.2 [15] for database storage.

The Same-CAT program, an open system for CAT implementation, created by Ricarte [20] during his Master's program at ICMC-USP, was used as the basis to construct the TAI-PI. The Same-CAT is a program that is able to implement of Computerized Adaptive Testing using items from

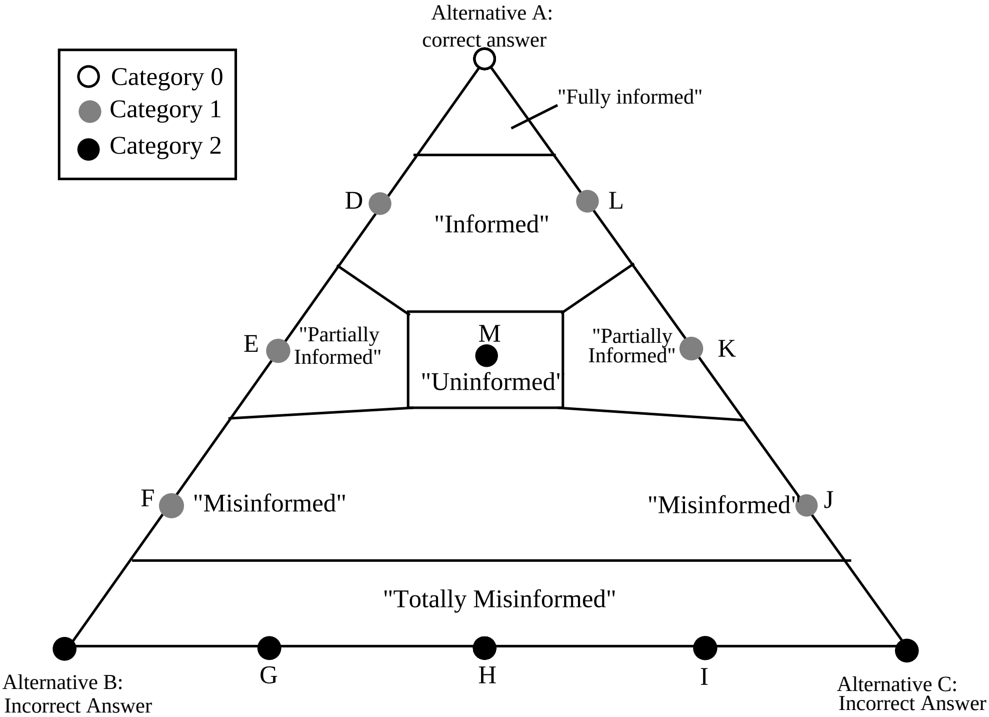

Figure 2: Equilateral triangle of APM with the characterization of the alternatives to the answers for each item of the test assuming that alternative A is correct and grouping of response categories to items adopted for TAI-PI.

the bank calibrated according to Samejima's Graded Response Model of the IRT [22], Kullback-Leibler criterion for item selection, the EAP method for latent trait estimation [3] and the Shadow Test Approach [31] for test constraints (content and test length, for example). The following subsections describe each of these methods, as well as the definition of the number of categories for Samejima's model and the starting and stopping criteria considered for the test.

3.1 Number of categories

The possibilities of answering each item of the EPI can be considered ordinal with 13 possibilities of answers grouped into 6 categories: fully informed, informed, partially informed, misinformed, totally misinformed and uninformed according to figure 2. As the purpose of this study is to transform EPI into a CAT, maintaining its history as much as possible, the GRM proposed by Samejima (1969) was initially adopted considering 6 categories for the TAI-PI. However, a descriptive and inferential analysis of this number of categories (frequencies of response in each category and the estimate item characteristic curve, respectively) showed that this number should be reduced and it was rearranged for three possible answers for each item.

This decision was based on the analysis of a subset of 74 items from the total database (with 167 items) that had at least 80 respondents from 2002 to 2013 (to achieve relatively reliable estimates of item parameters), and were presented in tests with anchor items (to ensure simultaneous equating of parameter estimates). The three categories defined in the adopted model are described in figure 2.

3.2 Samejima's model

Suppose $k = 0,1,2$ denoting the 3 categories in figure 2 arranged in ascending order, i.e., the higher the value of k, the closer it will be to the fully correct answer. The probability of an individual with proficiency θ choosing category k or anything greater than it in item i is given by:

$$P_{i,k}^{+}(\theta) = \frac{1}{1+exp[-a_i(\theta - b_{i,k})]}, \tag{3.1}$$

where $b_{i,k}$ denotes the difficulty parameter of category k of item i, a_i is the item discrimination parameter (equals for all categories of the item), and $P_{i,0}^{+} = 1$. In addition $b_{i,0} \leq b_{i,1} \leq b_{i,2}$.

The figure 3 shows the graph of the cumulative probability (3.1) with three item response categories.

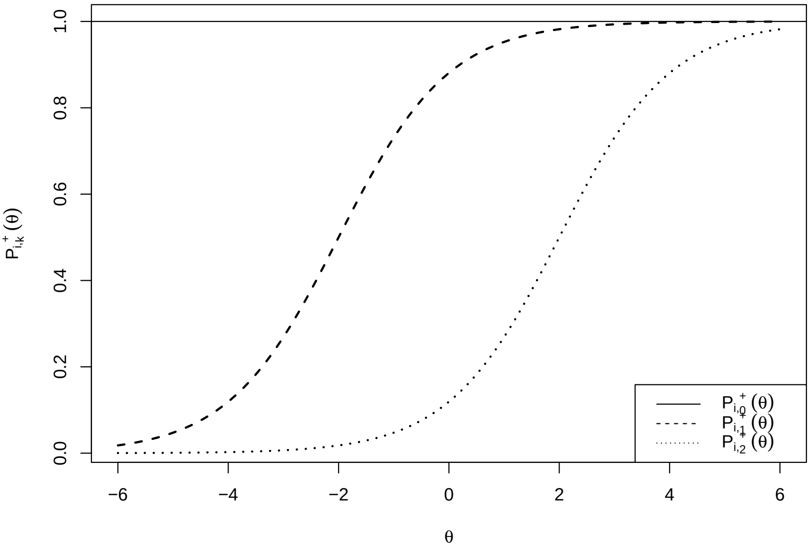

Figure 3: Example of a cumulative distribution curve in the Samejima's model with three item response categories with $a = 1$,b_1=-2 and b_2=2.

The probability of an individual with proficiency θ choosing the category k for item i is:

$$P_{i,k}(\theta) = P^{+}_{i,k}(\theta) - P^{+}_{i,k+1}(\theta), \tag{3.2}$$

where $P^{+}_{i,3}(\theta) = 0$ by definition. In figure 4, there is an example of the category response probability in Samejima's GRM for three categories.

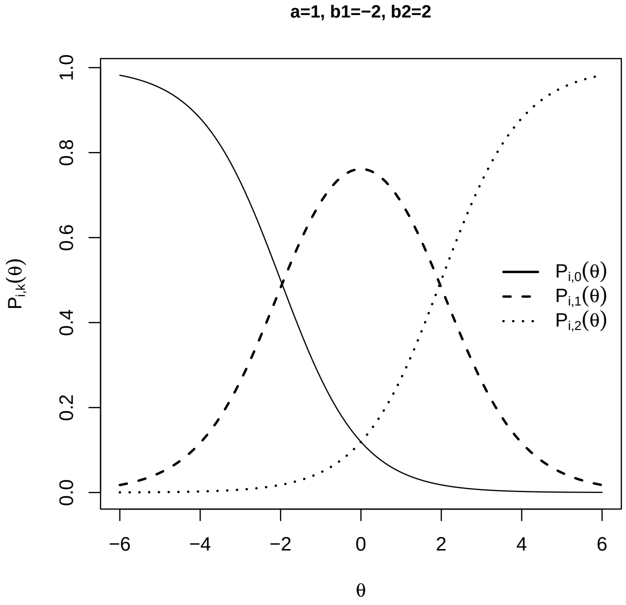

Figure 4: Example of the category response probability in Samejima's model with three item response categories.

3.3 Latent trait estimation

Assuming $l-1$ items already presented to the examinee and local independence, the likelihood function of model (3.2) is:

$$L(\theta \mid x) = \prod_{i=1}^{l-1} f(x_i \mid \theta) = \prod_{i=1}^{l-1}\prod_{k=0}^{2} P^{x_{ik}}_{i,k}(\theta), \tag{3.3}$$

where $x = (x_1, ..., x_{l-1})', x_i = (x_{i0}, x_{i1}, x_{i2})'$, and x_{ik} are variables that take value 1 if the k category of item i is chosen and 0, otherwise.

The EAP method was used for the estimation of θ following the well-known fact that Bayesian estimation works better in CAT than maximum likelihood methods, especially in early stages [31]. Baker [3], Baker and Kim [4], Mislevy and Stocking [19] suggested that the initial estimate of θ is obtained by EAP because the method can be calculated without the need for iterative methods, which reduces the time. The distribution, mean and variance a posteriori of θ, conditioned to the observed data, are given respectively by the expressions (3.4) and (3.5).

$$\pi(\theta) \propto L(\theta \mid x)g(\theta \mid \lambda), \quad \text{and} \quad E(\theta \mid x,\lambda) = \frac{\int_{\mathbb{R}} \theta L(\theta \mid x)g(\theta \mid \lambda)d\theta}{\int_{\mathbb{R}} L(\theta \mid x)g(\theta \mid \lambda)d\theta}, \tag{3.4}$$

$$Var(\theta \mid x,\lambda) = \frac{\int_{\mathbb{R}}(\theta - E(\theta))^2 L(\theta \mid x)g(\theta \mid \lambda)d\theta}{\int_{\mathbb{R}} L(\theta \mid x)g(\theta \mid \lambda)d\theta}, \tag{3.5}$$

where $g(\theta \mid \lambda)$ is the θ prior distribution and λ is the hyperparameter vector.

The numerical method of Gaussian quadrature [9] was used to calculate integrals numerically.

3.4 Item selection

The traditional item selection criteria based on the maximization of the Fisher information function was not adopted in TAI-PI. This is because if the current estimate of θ is far from its real value, which is very probable especially in the initial steps of a CAT, this criterion may be inappropriate. Chang and Ying [7] proposed an item selection procedure based on average global information, called the Kullback-Leibler (KL) item selection rule. It is based on the distance between the true ability θ and the current expected posterior ability estimate $\widehat{\theta}$ (EAP). The higher the value of this information, the greater the discrepancy between the two functions. For Samejima's model, it is given by:

$$K_i(\theta,\widehat{\theta}) = E\left[\log \frac{f(X_i|\theta)}{f(X_i|\widehat{\theta})}\right] = \sum_{k=0}^{2} P_{i,k}(\theta)\log \frac{P_{i,k}(\theta)}{P_{i,k}\cdot(\widehat{\theta})} \tag{3.6}$$

Assuming conditional independence among the responses, the KL after *l-1* administered items is written as:

$$K_{l-1}(\theta,\widehat{\theta}) = \sum_{i=1}^{l-1} K_i(\theta,\widehat{\theta}). \tag{3.7}$$

Because the true ability θ is unknown, Chang and Ying [7] proposed integrating (3.6) over a confidence interval for θ, $[\widehat{\theta} - \delta_l, \widehat{\theta} + \delta_l]$, and δ_l a decreasing function with relation to l. Thus, the criterion to select the next item to be shown in the test is given by:

$$i_l \equiv arg\max_i \left\{ \int_{\widehat{\theta}-\delta_l}^{\widehat{\theta}+\delta_l} K_i(\theta,\widehat{\theta})d\theta : i\varepsilon L \right\}, \tag{3.8}$$

where $\delta_l = z_\gamma/\sqrt{l}$ and L is the set of items not yet presented to the individual in the test.

3.5 Shadow test approach

The selection criteria expressed in (3.6) does not take into consideration any subjective restriction to the test composition. For example, a proper assessment of English proficiency should include items associated with each of the three modules: scientific text structure, reading comprehension and grammar conventions of language. The non-imposition of this restriction can generate tests with too many items of one module and no items for one (or both) of the others, distorting the desired assessment. Another practical aspect that must be considered is the number of items and texts in the test. Two individuals which perform tests with very different item numbers and/or numbers of texts can bring about a sense of injustice when comparing the estimates of the respective abilities. A person who answered a test with many texts or texts that are too long can complain that he/she was more tired than the one who answered a short test with few items, and the latent trait estimates can be incomparable. On the other hand, the examinee who answers the shorter test can complain that there was no ”time” (i.e., number of items) to demonstrate his/her actual proficiency.

One of the most interesting proposals in the literature to include these restrictions easily is the Shadow Test Approach [29].

The integer linear programming optimization method is used to obtain a subset (of size n previously defined) of the item bank that maximizes the information (3.7), taking $l-1=n$, subject to the desirable restrictions, also previously defined. A set of n items is obtained in each step of the algorithm, generating the test solution in that step. Importantly, one of the test restrictions in step i is that all items previously administered are present in the current test.

The item to be administered in step i test is the one not presented to the examinee in the previous steps, which belongs to the current test solution, and the one which presents the maximum information (3.6) for the current estimate latent trait value. After the selection, the unused items in the test return to the item bank as items available to be administered in the next steps of the Shadow test.

Special care should be taken when there are items associated with the same text or picture (generically called “stimulus”). In real situations, the ideal situation is to present all items associated with the same stimulus consecutively [30, 31]. In this work, when a new stimulus appears in a certain iteration of the Shadow test, all the items associated with that stimulus in the Shadow test of this iteration will be presented consecutively to the individual.

3.6 Starting and stopping criteria

The initial level of the latent trait is necessary for the test to start and it is related to the level of difficulty of the first selected item. [8] believes that the level of difficulty of the first item must be chosen so as to enable a reduction in the time of the test. Sukamolson [27] points out that usually the first item should have a level of medium difficulty when no previous information is available about the individual proficiency. As it is the case in this work, it was decided to start with an average $\widehat{\theta}_0 = 0$ for all the examinees.

The test stop criterion defines the moment that no more items need to be answered by the examinee. There are two criteria commonly used described in the literature: the first is to define a fixed number of items and the other is to define a minimum accuracy for the latent trait estimate, i.e., a predetermined minimum value for its standard error. This work considered a test with a fixed number of items. This is required by the Shadow test and is quite consistent with the EPI, as the students have a maximum of two hours to finish the exam and to avoid complains that could arise of different number of items have been responded among examinees.

4 SIMULATION STUDY

This simulation study was performed in software R and it was designed to illustrate the operation of a CAT in terms of latent trait estimation efficiency. Twelve scenarios were simulated combining 3 prior distributions in the EAP latent trait estimation method with 4 different numbers of items in the test (10, 20, 25 and 30). The bank consisted of 500 items with 3 categories of answers, generated using $a \sim \text{lognormal}(0.7, 0.1)$, in order to produce good values of discrimination parameters [18], and $b_i \sim N(0, 1.2)$, where $b_1 < b_2$, based on the Samejima's model with 3 categories of response. The EAP method was used to estimate the latent trait of the same model and Kullback-Leibler criterion was used as the item selection criterion. The start criterion of CAT was $\hat{\theta}_0 = 0$ for all individuals, and a fixed number of items in the test (10, 20, 25 and 30 items) was the assumed stop criterion.

Seven latent trait values were fixed as $\theta \in \{-3, -2, -1, 0, 1, 2, 3\}$ and 200 adaptive tests were simulated for each of these values. Latent trait estimates were obtained using the EAP method and the following three distributions: N(0,1), which is usually adopted in the literature, N(0,2) to provide less information about the parameter concerned and $N(\hat{\theta}_{l-1}, \sigma^2(\hat{\theta}_{l-1}))$ to consider the current estimate as a prior information of the individual proficiency. This work is the first to study the use of this approach as prior for the ability in CAT. Thus, there is a total of $7 \times 3 \times 200 = 4,200$ simulated tests.

To evaluate the quality of the recovery, it is necessary to include new indexes to represent the different individuals (j) and the different values for the fixed latent trait (k). The following measurements were obtained:

$$Bias = \sum_{j=1}^{200} \frac{\hat{\theta}_{jk} - \theta_k}{200} \quad \text{and} \quad MSE = \frac{\sum_{j=1}^{200} (\hat{\theta}_{jk} - \theta_k)^2}{200},$$

where $j = 1, \ldots, 200$ is the repetitions of each fixed value of θ, $k = 1, \ldots, 7$ indexes the 7 values set for θ, $\theta_k \in \{-3, -1, -2, 0, 1, 2, 3\}$ are the values set for θ, $\hat{\theta}_{jk}$ is the estimated θ_k obtained by the EAP method in the simulation after all test items (l=10, 20, 25 or 30) were answered.

The results are shown in tables 1 and 2. It can be observed that under all conditions that longer tests produce smaller Mean Square Errors (MSE) and an average bias very close to zero, especially when latent trait values dawn closer to the mean (0). In addition, increasing the size of the test from 25 to 30 items show very few improvement and even 20 items may be considered good enough on a test [31]. Except for $N(\hat{\theta}_{l-1}, \sigma^2(\hat{\theta}_{l-1}))$ prior distribution and for N(0,1) in cases when the latent traits are far from the average (i.e., near -3 and 3), the bias and mean square errors are small.The latter could be expected because of the generation scheme of the item bank ($b_i \sim N(0,1.2)$ causes low information for extreme values in the scale), the low prior variance assumed for θ combined with an insufficient number of items in the test to achieve efficiency in the estimation process. It is important to note how the prior distribution can affect the quality of the estimates when the aim is to reduce the size of the test, as in CAT.

Table 1: Bias of the estimators for different priors.

Prior	Test	True values θ						
	items	-3	-2	-1	0	1	2	3
N(0,1)	10	.3978	.1478	.0636	.0057	-0.0707	-0.1050	-0.6543
	20	.1756	.0768	.0373	.0180	-0.0423	-0.0246	-0.5341
	25	.1403	.0451	.0175	.0104	-0.0256	-0.0112	-0.5029
	30	.1200	.0497	.0168	.0045	-0.0154	-0.0093	-0.4754
N(0,2)	10	.1020	.0569	-0.0270	-0.0149	-0.0199	.1536	-0.1768
	20	.0467	.0222	-0.0144	-0.0110	-0.0032	.1192	-0.0814
	25	.0482	.0237	-0.0086	-0.0134	-0.0043	.1064	-0.0636
	30	.0486	.0241	-0.0100	-0.0099	-0.0037	.1153	-0.0558
$N(\hat{\theta}_{i-1}, \sigma^2(\hat{\theta}_{i-1}))$	10	.5851	.2588	.0557	-0.0056	-0.0599	-0.1677	-0.7912
	20	.3225	.1229	.0404	-0.0010	-0.0311	-0.0718	-0.6824
	25	.2759	.0971	.0228	-0.0038	-0.0337	-0.0550	-0.6565
	30	.2391	.0820	.0210	-0.0014	-0.0287	-0.0536	-0.6352

Considering only the test sizes of 20, 25 and 30 items, figure 5 shows the plots of the true latent trait values in function of the estimated ones. It can be seen that for all considered scenarios, the estimates are very close to the real values. This fact will help in choosing the number of items to be presented in the test.

5 APPLICATION

The same subset of the item bank with 74 items, described in Section 3.1, was used for the real application. An additional filter of these 74 items was carried out in order to select items with a discrimination parameter above 0.7 and testlets with at least 3 items (or items with no texts associated) resulting in 40 items used for the computerized adaptive test application. Item parameter estimates were obtained in the metric (0,1), i.e. considering mean equals to 0 and variance equals to 1 for the latent trait, using the collected responses of EPI applications from 2002 until 2013.

Table 2: Mean square error of the estimators for different priors.

Prior	Test	True values θ						
	items	-3	-2	-1	0	1	2	3
N(0,1)	10	0.2498	0.1295	0.0760	0.0735	0.0790	0.1015	0.4478
	20	0.0795	0.0525	0.0430	0.0395	0.0427	0.0662	0.3105
	25	0.0568	0.0397	0.0355	0.0314	0.0352	0.0663	0.2800
	30	0.0463	0.0397	0.0325	0.0249	0.0304	0.0608	0.2559
N(0,2)	10	0.0899	0.1164	0.0896	0.0778	0.0962	0.2381	0.0923
	20	0.0516	0.0580	0.0550	0.0438	0.0400	0.1551	0.0675
	25	0.0426	0.0462	0.0492	0.0339	0.0321	0.1350	0.0667
	30	0.0371	0.0394	0.0395	0.0287	0.0308	0.1391	0.0730
$N(\hat{\theta}_{i-1}, \sigma^2(\hat{\theta}_{i-1}))$	10	0.4508	0.1217	0.0872	0.1041	0.0764	0.0920	0.6352
	20	0.1641	0.0539	0.0433	0.0487	0.0416	0.0502	0.4755
	25	0.1251	0.0446	0.0392	0.0358	0.0356	0.0474	0.4408
	30	0.0953	0.0427	0.0339	0.0305	0.0280	0.0473	0.4133

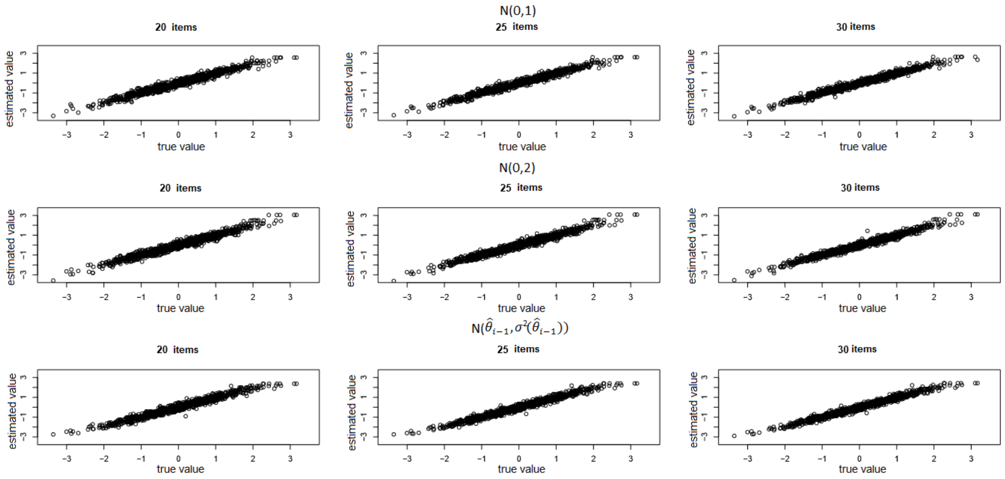

Figure 5: True values versus estimated values of θ for the simulations considering different priors and test sizes.

The real application of TAI-PI occurred in May 2014 for 59 graduate students from the CCMC and PIPGES graduate programs. As the IRT assessment method proposed in TAI-PI had not been validated yet and no cut-off point on the latent trait scale was studied, it was decided that the test should be developed with all the 40 items previously filtered. Note that the APM method was be used for student classification and it was important to ensure fairness in this classification. In the real aplication, the first 25 items presented to the individuals were selected according to the CAT methodology described in the previous section (Kullback-Leibler maximization of information, EAP method for latent trait and so on). The subsequent 15 items were presented following the item bank sequence in order to make sure that all students answered the same items (although with a different order of presentation). At this point, a validation study is necessary before adopting the new classification criteria to the students. The division of the test into 25 items presented in an adaptive way and 15 ones in the sequence to complete the bank was determined due to the simulations results, in which a 25-item test showed to be long enough for a good recovery. Therefore, the application served as a pilot study of CAT implementation and evaluated the student's English proficiency fairly. The structure of the 40-item bank application is shown in table 3.

Table 3: Item bank of TAI-PI, May 2014.

Module	Stimulus	Number of Items
1	Abstract 1	3
	Abstract 2	4
	Abstract 3	3
	Abstract 4	4
2	Introduction 5	7
	Introduction 6	4
3	none	15

To perform the adaptive part of the test (selecting the first 25 items), the following content restrictions were implemented via the Shadow test [29]:

- Exactly three texts from Module 1 (with at least 3 and maximum 4 items each text)
- Exactly one text from Module 2 (with at least 4 and maximum 7 items)
- A maximum of 11 items from Module 3
- Maximum of 2015 words contained in the texts covering the 25 items

The stored data for each examinee corresponds to the answers to all items of the bank, the latent trait estimate and respective standard error in each of the 25 steps of CAT and also after 40 items from the bank were answered, as well as the order of the item presentation. Table 4 shows the latent trait estimates of the students after the response to all the 40 items of the bank, as well as the classification (pass or fail) based on the APM methodology (considering all the 40 item answers). As can be seen highlighted in gray, 4 students have low θ estimates and passed using APM. Note that table 5, the results of these individuals correspond to the minimum (except for in the case of the 17th student) for to pass by the APM.

Figure 6 shows the boxplot of the estimated proficiencies in TAI-PI for the 59 students classified as pass or fail, according to the APM. It can be seen that: (i) a possible cut-off value for classification of the student in the latent trait scale may be defined around -0.5, and (ii) there is little overlap between the latent trait distributions of groups "pass" and "fail", especially after 40 answered items (as was expected, as there is more information in it than in the 25-item test).

Table 4: Estimates of latent traits (after 25 and 40 items answered) and classification by the APM in the TAI-PI application May 2014.

Número	Indivíduo	Theta25	Theta40	MPA
1	10	-2,07	-1,68	FAILED
2	1	-1,62	-1,66	FAILED
3	6	-1,29	-1,60	FAILED
4	35	-1,39	-1,59	FAILED
5	37	-1,-19	-1,44	FAILED
6	26	-1,34	-1,41	PASSED
7	52	-1,09	-1,37	FAILED
⋮	⋮	⋮	⋮	⋮
12	12	-0,96	-1,11	FAILED
13	38	-0,92	-1,08	PASSED
14	27	-1,01	-0,99	FAILED
15	3	-0,28	-0,76	FAILED
16	54	-0,29	-0,75	FAILED
17	22	-0,58	-0,73	FAILED
18	47	-0,34	-0,69	PASSED
19	49	-0,55	-0,69	FAILED
⋮	⋮	⋮	⋮	⋮
28	19	-0,17	-0,52	FAILED
29	17	-0,56	-0,48	PASSED
30	40	-0,28	-0,41	FAILED
31	53	-0,58	-0,41	FAILED
32	44	-0,32	-0,40	PASSED
⋮	⋮	⋮	⋮	⋮
59	46	1,17	1,08	PASSED

Table 5: Percentage of responses in each APM category for 4 students in the application of May 2014.

Student	%fully informed	%Informed	%Part. Informed	%Fully misinformed
26	0%	0%	90 %	10%
38	50%	0%	0%	25%
47	50%	12.5%	5%	25%
17	52.5%	5%	7.5%	25%

In figure 7, the estimated proficiency and its standard error of a student at each step of the adaptive test is shown. It can be seen that when the student answers the item incorrectly (represented by

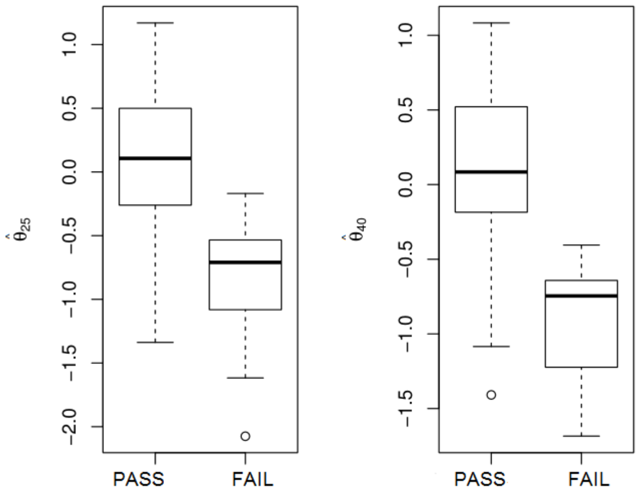

Figure 6: Proficiency boxplot of latent trait estimates after 25 and 40 answered items of 59 students categorized in pass or fail by APM criterion.

category 0), the subsequent estimated proficiency is a lower value. While, when the student answers the item correctly (category 2), the estimated proficiency increases. For partially correct answers (category 1), both situations can occur.

The purpose of EPI is not to estimate student proficiencies, but to classify them into “pas” or “fail”. A cut-off point in the latent trait scale is required for this aim. Thus, some studies were conducted using the data from the real application.

To contrast the IRT results with the APM results, which are the only available methods to evaluate the proficiency of theses students, the cutoff point definition was based on $\hat{\theta}_{40}$ (the latent trait estimate calculated using the 40 item responses). However, to measure the quality and efficiency of this cutoff point, studies were conducted based on the $\hat{\theta}_{25}$.

In figure 8 the ROC (Receiver Operation Characteristic) curve is presented considering various cutoff points based on $\hat{\theta}_{40}$ and assuming APM criteria as gold standard.

The cost-benefit method ([16], [17]) was used to find the optimal cutoff point, which is obtained when the slope of the ROC curve is given by:

$$S = \frac{1-p}{p} CR = \frac{1-p}{p} \frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}} \tag{5.1}$$

where p is the prevalence of “fail” in the test, C_{FP} is the cost of false positive, C_{TN} is cost of true negative, C_{FN} denotes the cost of false negative, and C_{TP} denotes the cost of true positive. Considering $CR = 1$, the value found for the cutoff point of the latent trait is -0.40. Figure 9

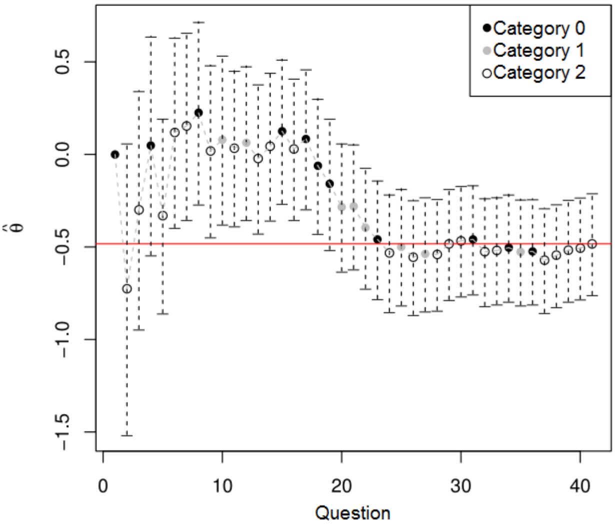

Figure 7: Proficiency estimated at each step of the TAI-PI for student 17. The horizontal gray line represents the final estimate after 40 answered items.

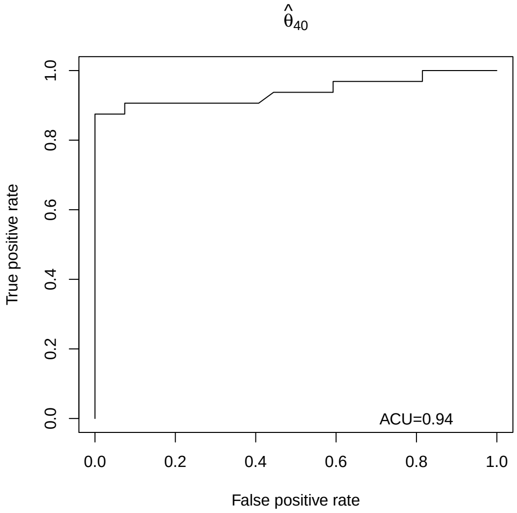

Figure 8: ROC curve for proficiency estimates ($\hat{\theta_{40}}$) considering APM as gold standard. ACU: area under curve.

shows the density estimates for “pass” and “fail” groups, in which you can be observed by the overlapping tails that there is little probability of making mistakes in the classifications.

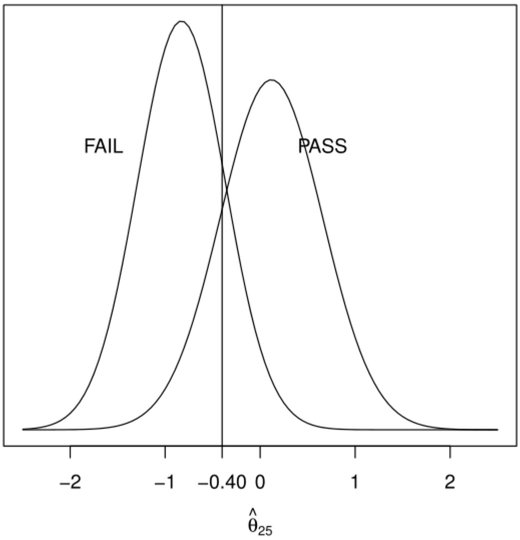

Figure 9: Density estimates for proficiency by group.

This classification error can be measured by the Expected Classification Accuracy method as proposed by [21]. Accuracy is an index to quantify the capacity of the test in reflecting true results, i.e., when the student is classified as “pas” by the cut-off point criterion and he/she passes the APM criterion, or the contrary , “fail” by the cutting point and “fail” the APM criteria. The expected accuracy is given by:

$$EA = \frac{\sum_{\theta_i<\theta_c} P(\hat{\theta} < \theta_c|\theta_i)}{n} + \frac{\sum_{\theta_i>\theta_c} P(\hat{\theta} > \theta_c|\theta_i)}{n} \tag{5.2}$$

where $\theta_c = -0.40$. The expected accuracy found was 0.85, i.e, it is expected to make a mistake in 15% of the classifications, a relatively small value considering the fact that the student usually has a second chance to pass the English test offered by ICMC for free or to apply of another (charged) exam, such as TOEFL, IELTS or CPE.

A simulation study was also performed to study the expected accuracy value for these proposed cut-off points in a more controlled set of proficiencies. Considering this, 1,000 proficiency values were taken from a sequence -1.4 to 0.6 with space of 0.002. Considering exactly the same scheme for the simulation presented in Section 4, adopting the prior distribution for individual proficiency as N(0,1) and a 25-item test, the accuracy obtained is shown in table 6. A misclassification percentage of approximately 6% corroborates the results reached in the real application.

Table 6: Accuracy for the cut-off point in the simulation study.

	$\theta < -0.4$	$\theta \geq -0.40$
$\hat{\theta} < -0.40$	0.4585	0.024
$\hat{\theta} \geq -0.40$	0.040	0.4755

6 CONCLUSION

The results of the simulations and the real application of TAI-PI indicate that the computerized adaptive test developed in this work assuming normal prior distribution for the latent trait at each step of the test, adopting the EAP estimation method and Shadow Test Approach is feasible to asses the language proficiency of English graduate students from ICMC. It enables the immediate compilation of the test results, it did not lead to difficulties in students performing the test, it discouraged cheating (one the items may differ among test) and it produced a low misclassification rate (when compared to the APM criterion). The results show that a 25-item test seems to be sufficient to satisfactorily estimate the latent trait, corroborating the finds of Vam der Linden and Pashley [31]. However, this number is not a consensus. Tseng [28], for instance, found significant differences in latent trait estimates between full bank test (180-item bank) and 30-item CAT in an assessment for English vocabulary size. Future research efforts should be dedicated to evaluate the influence of the characteristics of the items in the bank in the precison of latent parameters estimate. This work also shows that a cut-off point of -0,40 in the latent trait scale can be adopted in future assessments to classify the student as “pas” or “fail”. The TAI-PI is a modern and efficient form of assessment to be used for graduate students from ICMC and can be offered as a service to other institutions, as it is available online. It will be less time consuming for university lecturers and will provide a free way to evaluate the student’s English proficiency. The application considered simple constraints in Shadow test approach because it is a small scale application with a small item bank. Future studies should consider the enlargement of the number of items in the bank (which is already been done) and, consequently, more complex restrictions may be added for test design, such ones to avoid item overexposure.

RESUMO. Este trabalho descreve as etapas de transformação de um exame de proficiência em inglês acadêmico, aplicado via lápis-e-papel, em um teste adaptativo informatizado (TAI-PI) baseado em um modelo da Teoria de Resposta ao Item (TRI). O exame é composto por itens de múltipla escolha administrados segundo o método de Medida de Probabilidade Adminssível e é adotado no programa de pós-graduação do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP). Apesar do programa aceitar diversos exames que atestam a proficiência em inglês para indivíduos não-nativos de abrangência e reconhecimento internacionais, como o TOEFL (Test of English as a Foreign Language), IELTS (International English Language Testing System) e CPE (Certificate of Proficiency in English), por exemplo, a sua obrigatoriedade é incompatível com a forma de funcionamento da universidade pública do Brasil devido ao custo que varia de 200 a 300 dólares por exame. O software TAI-PI (Teste Adaptativo Informatizado para Proficiência em Inglês), que foi desenvolvido em Java e SQLite, será utilizado para a avaliação da proficiência em inglês dos alunos do programa desde o segundo semestre de 2013, de forma gratuita. A metodologia estatística implementada foi definida considerando a história e objetivos do exame e adotou o modelo de resposta gradual unidimensional de Samejima, o critério de Kullback-Leibler para seleção de itens, o método de estimação da esperança a posteriori para os traços latentes e a abordagem Shadow test para imposição de restrições (de conteúdo e tamanho da prova) na composição do teste de cada indivíduo. Uma descrição da estrutura do exame, dos métodos empregados, dos resultados das aplicações do TAI-PI a alunos de pós-graduação do ICMC e estudos de classificação dos alunos em aprovados e reprovados, são apresentados neste trabalho, evidenciando a boa qualidade da nova proposta adotada e aprimoramento do exame com a utilização dos métodos de TRI e TAI.

Palavras-chave: teste adaptativo computadorizado, teoria de resposta ao item, *shadow test*.

REFERENCES

[1] S. Aluísio, V. de Aquino, R. Pizzirani & O. de Oliveira. Assessing High-Order Skills with Partial Knowledge Evaluation: Lessons Learned from Using a Computer-based Proficiency Test of English for Academic Purposes. *Journal of Information Technology Education: Research*, **2**(1) (2003), 185–201.

[2] E.C. Aybek & R.N. Demirtasli. Computerized Adaptive Test (CAT) Applications and Item Response Theory Models for Polytomous Items. *International Journal of Research in Education and Science*, **3**(2) (2017), 475–487.

[3] F. Baker. "The basic of Item Response Theory. EUA: ERIC Clearinghouse on Assessment and Evaluation". , 2 ed. (2001).

[4] F.B. Baker & S.H. Kim. "Item response theory: Parameter estimation techniques". CRC Press (2004).

[5] B. Bloom. "Taxonomy of Educational Objectives". New York, Longman (1984).

[6] B. Bridgeman & F. Cline. Variations in Mean Response Times for Questions on the Computer-Adaptive GRE® General Test: implications for fair assessment. *ETS Research Report Series*, **2000**(1) (2000), i–29.

[7] H.H. Chang & Z. Ying. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, **20**(3) (1996), 213–229.

[8] R. Conejo, E. Millán, J.L. Perez-de-la Cruz & M. Trella. Modelado del alumno: un enfoque bayesiano. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, **5**(12) (2001), 50–58.

[9] M. Dehghan, M. Masjed-Jamei & M. Eslahchi. On numerical improvement of open Newton–Cotes quadrature rules. *Applied mathematics and computation*, **175**(1) (2006), 618–627.

[10] D. Eignor, C. Taylor, I. Kirsch & J. Jamieson. Development of a scale for assessing the level of computer familiarity of TOEFL examinees. *ETS Research Report Series*, **1998**(1) (1998), i–32.

[11] D.R. Eignor. Deriving Comparable Scores for Computer Adaptive and Conventional Tests: an example using the SAT1, 2. *ETS Research Report Series*, **1993**(2) (1993), i–16.

[12] I. Kirsch, J. Jamieson, C. Taylor & D. Eignor. Computer familiarity among TOEFL examinees. *ETS Research Report Series*, **1998**(1) (1998), i–23.

[13] A. Klinger. Experimental validation of learning accomplishment. In “Frontiers in Education Conference, 1997. 27th Annual Conference. Teaching and Learning in an Era of Change. Proceedings”, volume 3. IEEE (1997), pp. 1367–1372.

[14] C.B. Kreitzberg, M.L. Stocking & L. Swanson. Computerized adaptive testing: Principles and directions. *Computers & Education*, **2**(4) (1978), 319–329.

[15] R.F. Lans. Introduction to SQL: mastering the relational database language. (2006).

[16] B.J. McNeil, E. Keeler & S.J. Adelstein. Primer on certain elements of medical decision making. *New England Journal of Medicine*, **293**(5) (1975), 211–215.

[17] C.E. Metz. Basic principles of ROC analysis. In “Seminars in nuclear medicine”, volume 8. Elsevier (1978), pp. 283–298.

[18] R.J. Mislevy. Bayes modal estimation in item response models. *Psychometrika*, **51**(2) (1986), 177–195.

[19] R.J. Mislevy & M.L. Stocking. A consumer’s guide to LOGIST and BILOG. *Applied psychological measurement*, **13**(1) (1989), 57–75.

[20] T.A.M. Ricarte. “Teste adaptativo computadorizado nas avaliações educacionais e psicológicas”. Ph.D. thesis, Universidade de São Paulo (2013).

[21] L.M. Rudner. Expected classification accuracy. *Practical Assessment, Research & Evaluation*, **10**(13) (2005), 1–4.

[22] F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, (1969).

[23] W.A. Sands, B.K. Waters & J.R. McBride. “Computerized adaptive testing: From inquiry to operation.”. American Psychological Association (1997).

[24] D. Segall. Score equating verification analyses of the CAT-ASVAB. *Briefing presented to the Defense Advisory Committee on Military Personnel Testing. Williamsburg, VA, USA*, (1993).

[25] E.H. Shuford Jr, A. Albert & H.E. Massengill. Admissible probability measurement procedures. *Psychometrika*, **31**(2) (1966), 125–145.

[26] D. Spenassato, A.C. Trierweiller, D.F. de Andrade & A.C. Bornia. Computerized Adaptive Testing Applied in Educational Evaluations. *Brazilian Journal of Computers in Education*, **24**(02) (2016), 1.

[27] S. Sukamolson. Computerized test/item banking and computerized adaptive testing for teachers and lecturers. *Information Technology and Universities in Asia–ITUA*, (2002).

[28] W.T. Tseng. Measuring English vocabulary size via computerized adaptive testing. *Computers & Education*, **97** (2016), 69–85.

[29] W.J. Van der Linden. Constrained adaptive testing with shadow tests. *Computerized adaptive testing: Theory and practice*, (2000), 27–52.

[30] W.J. Van der Linden & C.A. Glas. “Elements of adaptive testing”. Springer (2010).

[31] W.J. Van der Linden & P.J. Pashley. Item selection and ability estimation in adaptive testing. In “Elements of adaptive testing”. Springer (2010), pp. 3–30.

[32] H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green & R.J. Mislevy. “Computerized adaptive testing: A primer”. Routledge (2000).

[33] H. Wainer & X. Wang. Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, **37**(3) (2000), 203–220.

[34] D.J. Weiss. Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, **6**(4) (1982), 473–492.