Journal of the Brazilian Computer Society

RESEARCH Open Access

ALICAT: a customized approach to item selection process in computerized adaptive testing

Victor M. G. Jatobá[1], Jorge S. Farias[2], Valdinei Freire[1], André S. Ruela[1*] and Karina V. Delgado[1]

Abstract

Computerized adaptive testing (CAT) based on item response theory allows more accurate assessments with fewer questions than the classic paper and pencil (P&P) test. Nonetheless, the CAT construction involves some key questions that, when done properly, can further improve the accuracy and efficiency in estimating the examinees' abilities. One of the main questions is in regard to choosing the item selection rule (ISR). The classic CAT makes exclusive use of one ISR. However, these rules have differences depending on the examinees' ability level and on the CAT stage. Thus, the objective of this work is to reduce the dichotomous test size which is inserted in a classic CAT with no significant loss of accuracy in the estimation of the examinee's ability level. For this purpose, we analyze the ISR performance and then build a personalized item selection process in CAT considering the use of more than one rule. The case study in *Mathematics and its Technologies* test of the ENEM 2012 shows that the Kullback-Leibler information with a posterior distribution (*KLP*) has better performance in the examinees' ability estimation when compared with Fisher information (*F*), Kullback-Leibler information (*KL*), maximum likelihood weighted information (*MLWI*), and maximum posterior weighted information (*MPWI*) rules. Previous results in the literature show that CAT using *KLP* was able to reduce this test size by 46.6% from the full size of 45 items with no significant loss of accuracy in estimating the examinees' ability level. In this work, we observe that the *F* and the *MLWI* rules performed better on early CAT stages to estimate examinees' proficiency level with extreme negative and positive values, respectively. With this information, we were able to reduce the same test by 53.3% using the personalized item selection process, called ALICAT, which includes the best rules working together.

Keywords: Computerized adaptive testing, Item response theory, Fisher information, Item selection rule, Item selection method

Introduction

The evaluation of students has always been very important in the learning process. In computer-assisted education, learning systems can generate tests that help students to identify if they have achieved an appropriate level of knowledge [1]. An example of such system is the computerized adaptive testing (CAT). CATs are computer-administered tests that efficiently reduce the number of items (questions) while maintaining a good estimation of the respondent's ability level [2]. In classic CAT, a question is initially selected and, for each new question, the student's ability level is estimated. If the stopping criterion is not met, another question is selected.

There are several ways to assess the examinees' ability level. A recently used model is the item response theory (IRT) [1, 3, 4]. This theory includes a set of mathematical models that attempts to determine the probability of an examinee to correctly answer a given question, considering the item's characteristics and the examinee's abilities [5]. This is the model adopted, for example, by the Brazilian High School-level exam, called *Exame Nacional do*

*Correspondence: andre.siqueira.ruela@gmail.com
[1] Department of Information Systems, School of Arts, Sciences and Humanities, University of São Paulo, Rua Arlindo Bettio n^o1000, São Paulo 03828-000 Brazil
Full list of author information is available at the end of the article

Ensino Médio (ENEM), to compute the students' overall performance [6].

The IRT-based CATs provide more accurate tests [2] because it is possible to identify the student's ability level and thus select a sequence of items that fits the user's knowledge [7]. CATs have several advantages over the paper and pencil (P&P) test format [8]. One is the possibility to accurately estimate the examinees' latent abilities, considering a reduced number of questions [9, 10].

However, the construction of CATs involves a number of key features, such as choosing the test initialization method, the stopping criteria, and the item selection rule (ISR) [9]. In addition, depending on the scenario, CATs can use, along with the ISRs, some mechanism to control item exposure, or some kind of content balancing [11]. An appropriate choice of such key features can improve accuracy and efficiency to estimate the examinee's abilities, especially in relation to the ISRs [11–14].

The item selection process considered in the classic CAT construction involves choosing only one ISR [15]. However, these rules have advantages and disadvantages depending on the examinees' ability level and the ongoing test stage, i.e., the number of questions that have already been selected.

This paper presents a novel approach for the CAT construction, denominated ALICAT (personALIzed CAT), which aims to customize the item selection process. This customization is done by dynamically choosing an ISR based on the student's performance and the actual stage of the test. The goal of this proposal is to reduce the length of dichotomous tests—which considers only correct and incorrect answers—that are embedded in a CAT environment, with no significant loss of accuracy in estimating the examinees' abilities. To achieve that goal, we analyze the performance of different ISRs to properly customize the selection process of items, considering the use of more than one ISR.

Background

In this section, we present a brief introduction of the theories used in this paper and the main concepts involving the configuration of a CAT. Finally, some details of the ENEM test are explained, justifying its use in the case study of this work.

Computerized adaptive testing

Computer-based adaptive tests, also known as test-generating systems, are computer-administered tests that efficiently reduce the number of items while maintaining a good diagnosis of the examinee's performance [2]. This tool emerged as an alternative to the conventional classical test theory (CTT) exams applied in paper and pencil format [16].

CAT systems are composed of five main components: bank of items (BI), initialization criteria, ISR, method to estimate the examinee's ability level, and stopping criteria [17]. Compared to conventional tests, the main advantages of CATs are (i) the generation of faster and shorter tests; (ii) the possibility of test application in flexible hours; (iii) since each student receives a unique set of items, it is much more difficult for the examinee to cheat; (iv) better control of the exposure of questions; (v) more accurate examination; (vi) the possibility of immediate feedback; and (vi) a reduction of the number of items without reducing the precision in estimating the examinee's ability level [1, 9, 18].

It should be noted that the key questions found in CAT are [9] as follows: (i) How to choose the first question given that, initially, nothing is known about the examinee? (ii) How to choose the next item after the response of the current item? (iii) How to choose the ideal time to stop the test?

These questions justify the existence of different models that support the construction of the CAT, which can meet each one in a particular way. Some examples of these models are the sequential probability ratio test [19], the combination of granularity hierarchies and Bayesian networks [20], measurement decision theory [21], and IRT. However, IRT is the most widely used model in CAT systems [1, 3, 4].

Item response theory

The IRT consists of mathematical models that establish the probability of an examinee to hit a given question, given the item's characteristics and the examinee's abilities [5]. One of the great advantages of the IRT is that the computation of the examinee's ability level and the item parameters (difficulty and discrimination) are independent of the sample of items [22]. This allows the comparison of populations if tests have some common items or the comparison between individuals from the same population that have been submitted to totally different tests.

In this paper, we employ the logistic model with three parameters (ML3) for calculating the probability of correct or incorrect response to dichotomous items. In this model, the probability of user j, with ability θ_j, to correctly answer the item i can be calculated by:

$$P_i\left(\theta_j\right) = P\left(U_{ji}=1|\theta_j\right) = c_i + (1-c_i)\,\frac{1}{1+\exp^{-D.a_i.(\theta_j-b_i)}}, \tag{1}$$

where

- U_{ji} is a variable that assumes the value 1 when the examinee j answers the item i correctly and 0 otherwise;

- b_i is the difficulty parameter of the question;
- a_i is the discriminating power that each question has to differentiate the examinees who master, from those who do not master, the evaluated ability in item i [6]. The value of a_i is proportional to the derivative of the curve tangent at the inflection point at point b_i. Low values of a_i indicate that the item has little power of discrimination, that is, students with quite different abilities have similar probabilities to hit the question. For very high values, students are separated into two groups: those with abilities below b_i and those above b_i;
- c_i is the guessing probability of item i, i.e., the probability that a participant will hit the correct answer randomly and not by mastering the required ability [6];
- θ_j is the latent ability of user j; and
- D is a constant scale factor. Usually, the value 1.7 is used so that the logistic function gives results similar to the normal function [23].

The item selection rule

Choosing an ISR directly influences the efficiency and accuracy in estimating the ability of CAT respondents compared to P&P tests [24]. However, whether in CAT or P&P, accuracy is low at the early stages, when few questions have been answered [12]. Hence, higher accuracy depends on choosing a proper item selection criterion [11–14].

There are two well-established general approaches for selecting items in CATs: (i) information-based and (ii) Bayesian [25]. The first selects the item that provides more information about the estimation of a specific ability level and the second selects items based on prior and posterior distributions of estimation of a specific ability level [26].

The Fisher information (*F*) [27] technique is an example of an information-based ISR. An alternative is the Kullback-Leibler information (*KL*) [28], which is also based on information and is a general measure for the *distance* between two distributions [29]. According to [12], the *F* rule is more accurate to estimate the ability level of average users in the early stages of CATs if compared to the maximum posterior weighted information (*MPWI*) [30] and Kullback-Leibler information with a post distribution (*KLP*) [28], which are Bayesian rules. On the other hand, according to [11], the *MPWI* and *KLP* strategies have better results than the *F* rule for extreme ability levels (very high or very low values).

There are also studies that indicate that the maximum likelihood weighted information (*MLWI*) [31] and *MPWI* rules present a better general performance, in comparison with the *F* rule [29–31]. These results were obtained in applications that are inserted in the context of dichotomous tests. In addition, these applications do not make use of item exposure controls or content balancing controls.

To build the proposed CAT, in this paper, we perform a systematic analysis of the different aforementioned ISRs. In this analysis, we aim to investigate whether the application of an ISR is more appropriate than another, considering the examinees' ability level and the test stage.

Ability estimation

There are many methods to estimate the examinees' ability level such as maximum likelihood estimator (MLE), expected a posteriori (EAP), Bayes model (BM), maximum a posteriori estimator (MAP), and weighted likelihood estimator (WL). Since in this work the method for the estimation of abilities was EAP, we explain this method next, but before we need to explain MLE.

The MLE [27] computes the examinee's ability estimation $\hat{\theta}$ that maximizes the following function:

$$L_I\left(\theta_j\right) = \prod_{i=1}^{I} P_i\left(\theta_j\right)^{U_{ji}} Q_i\left(\theta_j\right)^{1-U_{ji}}, \tag{2}$$

where P_i is the probability of user j to correctly answer item i (Eq. 1), Q_i is the probability of an incorrect answer of item i ($Q_i = 1 - P_i$), and I is the test length.

The EAP [32] estimator is an alternative estimator. This estimator computes the mode of the posterior mean:

$$\hat{\theta}_{\mathrm{EAP}} = \frac{\int_{-\infty}^{+\infty} \theta\, f\left(\theta_j\right) L_I\left(\theta_j\right) d\theta_j}{\int_{-\infty}^{+\infty} f\left(\theta_j\right) L_I\left(\theta_j\right) d\theta_j}, \tag{3}$$

where $f(\theta_j)$ is the prior distribution.

The National High School Exam

The National High School Exam (ENEM) is a P&P exam promoted by the Brazilian National Institute for Educational Studies and Research Anísio Teixeira (INEP) [33]. It was created in 1998, but was reformulated in 2009, where it was used as a selection criterion for entry to universities [34].

The ENEM uses the concept of competencies, which translates into skills, knowledge, and attitudes to solve each problem situation. The exam is structured in five tests. One of the tests is the writing test, and the others are structured in four macro-areas, which are Nature Sciences and their Technologies; Human Sciences and their Technologies; Languages, Codes and their Technologies; and Mathematics and its Technologies [35].

The test items are dichotomous with the possibility of multiple choice questions, and to calculate the participant's score, the exam began to adopt the IRT in 2009. The logistic model used by IRT is the three-parameter model

(ML3). Each of the four tests has 45 items, and the calculation of the respondents' ability level is estimated by the EAP method.

Following, we define our proposed CAT and, subsequently, we describe our research method.

The ALICAT approach

In the standard CAT approach, only one ISR is used to select the sequence of questions. However, depending on the respondent's ability and the ongoing stage of the test, the performance of the selection rules can vary considerably [11, 12, 29–31]. That said, our proposal is to personalize the item selection process, choosing an ISR dynamically, considering the respondent's performance in fulfilling the test and the current stage of the test.

The ALICAT approach chooses the proper selection rule based on the users' answer pattern. If a given user correctly answers the first **p** questions, then for the next **q** items the ALICAT selects the ISR with the best performance to estimate the scores for users with high-level abilities. On the other hand, if a user answers wrongly the first **p** questions, the ALICAT does the opposite. In this case, for the next **q** questions, it selects the ISR with the best performance for users with low-level abilities. In any other case, the selected rule is the one with the best average performance (see Fig. 1).

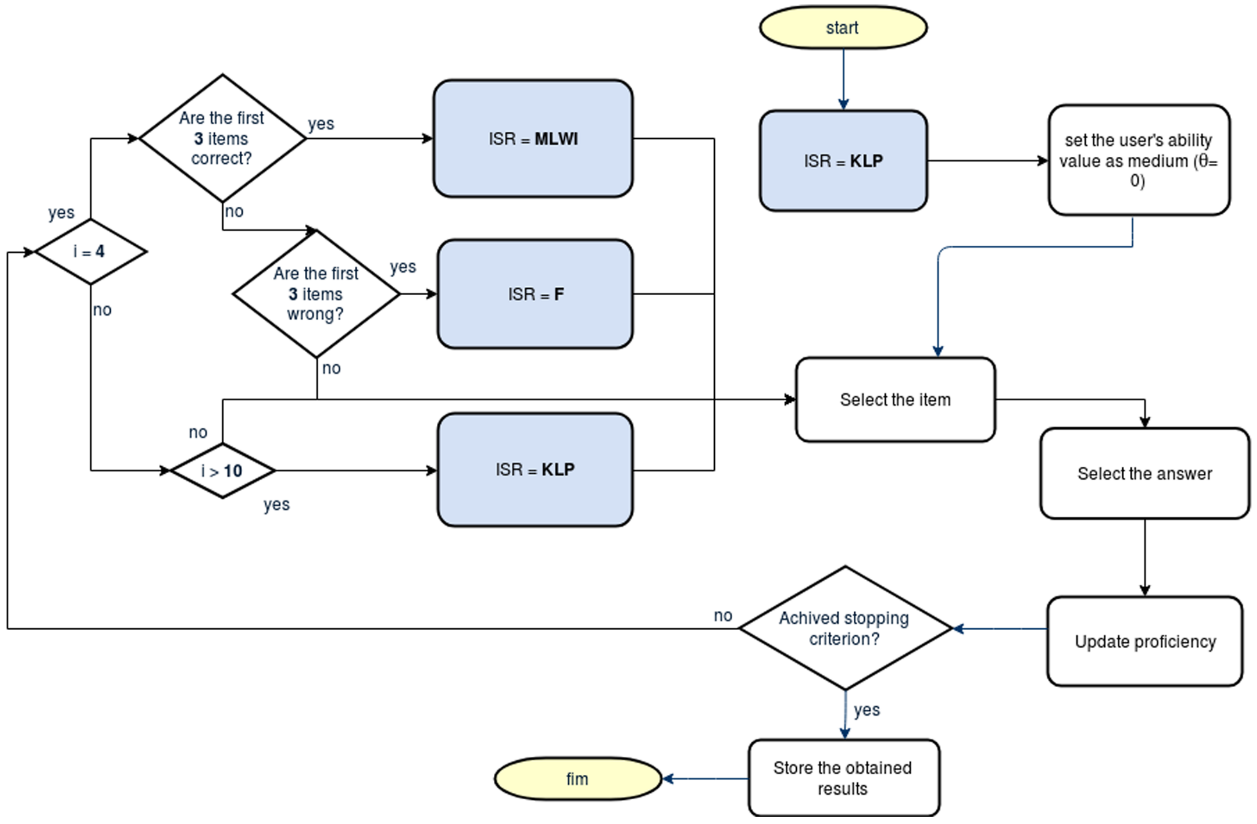
In Fig. 1, variable i represents the number of selected items. Variables p and q (with $p < q$) represent the test moment at which another ISR, different from the one that started the test, can be selected. In the figure, the squares in blue contain the test moment at which different rules are selected.

The strategy to re-use the best general rule after applying **q** items is due to the fact that, as the test grows, the performance of all rules tends to remain similar [12].

Consequently, considering a given test scenario, to execute the ALICAT approach, it requires identifying the ISR that presents the best average performance with all users, the best performance with users that have high ability level, and the best performance with users that have low ability level. Following, we describe how we defined these selection rules.

Research method

In this section, we describe the data and the item bank. Also, we elucidate the processes of conducting and evaluating the comparative study of the item selection rules.



start
ISR = KLP
set the user's ability value as medium (θ= 0)
Select the item
Select the answer
Update proficiency
Achived stopping criterion?
yes
Store the obtained results
fim
no
i = 4
yes
Are the first 3 items correct?
yes
ISR = MLWI
no
Are the first 3 items wrong?
yes
ISR = F
no
no
i > 10
no
yes
ISR = KLP

Fig. 1 Flowchart of the ALICAT approach

Finally, we present the configuration and evaluation form of the ALICAT approach.

Data and item bank

We used the data from the 2012 ENEM exam, which are public and have been taken from the transparency portal [36]. We choose this year in order to compare some of our results with the obtained results by Spenassato et al. [37].

The data model is composed of a set of randomly selected dichotomous responses. Each answer u can contain only two possible values, which are 1 for *correct* and 0 for *incorrect*, such that $u \in \{0, 1\}$, characterizing our distribution as a Bernoulli one.

As an assumption to be used throughout the text, we accept that our sample is independent and equally distributed, since the responses from different individuals are independent.

The initial sample comprised 1,000,000 respondents from the pink version of the *Mathematics and its Technologies* test. The values of a, b, and c parameters of the 45 items were taken from [37]. These parameters served as a basis for estimating the examinees' ability level.

Examinees' ability estimation

To estimate the examinees' ability level, we used the software ICL [38] and the method for the estimation of abilities was the expected a posteriori (EAP), the same one used in [37]. The scores estimated for the complete test of 45 items that correspond to the P&P test are called *true scores* and represented by θ. In time, let us define here $\hat{\theta}$ as the respondent's partial *estimated score*, which is assessed during the CAT execution.

Then, the computed true scores were ranked in 10 groups between − 2 and 3.5 (see Table 1). The first goal was to understand the behavior of the estimation step. The lowest and highest θ values were, respectively, − 1.716215 and 3.083216. It can be noted that there is a small number of respondents with high θs (greater than 2) and low θs (lesser than − 1.5).

Table 1 Sample size and lowest, highest, and average θ for each true score interval

True θ interval	Sample size	Lowest θ	Highest θ	Average θ
[− 2 ; − 1.5]	12193	− 1.716215	− 1.500013	− 1.58
[− 1.5 ; − 1]	121331	− 1.499992	− 1.000002	− 1.19
[− 1 ; − 0.5]	211557	− 0.999995	− 0.500002	− 0.75
[− 0.5 ; 0]	193117	− 0.499994	− 1e−06	− 0.25
[0 ; 0.5]	163131	1.3e−05	0.499995	0.24
[0.5 ; 1]	139804	0.500006	0.999997	0.74
[1 ; 1.5]	95364	1.000001	1.499987	1.23
[1.5 ; 2]	53275	1.500001	1.999989	1.71
[2 ; 2.5]	9435	2.000034	2.499647	2.16
[2.5 ; 3.5]	793	2.500007	3.083216	2.66

As in [37], 500 respondents were randomly selected from each group, thus obtaining another sample with 5000 respondents. This was important to ensure that all groups of respondent levels are part of the CAT simulation stage. From this group of 5000 respondents, we considered only those who answered at least 40 items, totaling 4979 respondents.

CAT configuration

For the assembly of CATs, we use the package `catR` [39] from the software R. There was no need to implement any criterion of item exposure because all evaluated respondents were submitted to the same 45 items. Also no restrictions for content balance has been developed as it is not disclosed which content each question belongs to.

To identify the length of the CATs, we applied the same methodology used by [37] that is described in the following. This methodology allows finding the number of questions necessary for a CAT to be able to estimate, with a certain degree of precision, the respondent's score. For this purpose, it is necessary to verify at which point of the test the accuracy of the $\hat{\theta}$ estimation remains stable.

To verify the stability point, we consider the calculation of the standard error (*SE*) using the *EAP* method that satisfies the following equation:

$$\mathrm{SE}_{I,j}\left(\hat{\theta}_{\mathrm{EAP}}\right) = \left[\frac{\int_{-\infty}^{+\infty}\left(\theta_j - \hat{\theta}_{\mathrm{EAP}}\right)^2 f\left(\theta_j\right)\, L_I\left(\theta_j\right)\, d\theta_j}{\int_{-\infty}^{+\infty} f\left(\theta_j\right)\, L_I\left(\theta_j\right)\, d\theta_j}\right]^{1/2}. \quad (4)$$

SE was estimated by the *semTheta* method available in the `catR`, an R package.

Thus, we compute $\mathrm{SE}_{I,j}$ that corresponds to the estimated standard error of the respondent j score, after applying I items. The stability point, for an examinee j, is the test moment, in which the difference of the standard error of the examinee score between the applied current item ($SE_{I,j}$) and the previous item ($SE_{I-1,j}$) is less than 1% of the standard error of his score taking into account the previous item. Thus, the stability point for the examinee j is the value of I that satisfies the following equation [37]:

$$\left|\mathrm{SE}_{I,j} - \mathrm{SE}_{I-1,j}\right| < \left|0,01 \times \mathrm{SE}_{I-1,j}\right|, \quad (5)$$

in which variable j represents the evaluated respondent, ranging from 1 to 4979. Variable I refers to the item, ranging from 1 to at least 40, and at most 45. When the stability point I is found, the test for examinee j can be completed with no considerable loss on the accuracy of the score estimations [37].

Figure 2 illustrates a hypothetical example of CAT application for a respondent with *true score* equal to 1.1 for a 45-item test. The value 1 (one) represents that the user gives a correct response and 0 (zero) represents a wrong response. The yellow line represents the value of the difficulty parameter of each sequentially selected item. We can observe that from item 13, the estimation of the current score ($\hat{\theta}$) is already very close to the true score (θ). This means that if the behavior of all the evaluated users was similar, value 13 could be a good length for the CAT. In the example, this would represent a reduction of 71.1% of the original 45-item test.

For each CAT, one of the item selection rules (*F*, *KL*, *KLP*, *MLWI*, and *MPWI*) were applied, with the initial $\hat{\theta}$ set to 0 (zero). In the CAT execution, the stability point of each respondent was identified. All points were organized in 10 groups, ranging from − 2 to 3.5. For each group, we compute the average of the points belonging to the group. Thus, the CAT length n was defined by the highest average in the 10 groups.

Evaluation of the ISRs

We follow the same method proposed in [37] and [12] to evaluate the ISRs. After computing the maximum length n of the test for each ISR, they are executed again, now considering the new fixed length.

To evaluate the performance of the ISRs in the estimation of abilities, we computed the average bias (BIAS), defined in Eq. 6, and the root mean squared error (RMSE), defined in Eq. 7, of the score estimation:

$$\mathrm{BIAS}(n) = \frac{1}{R}\sum_{k=1}^{R}\left(\hat{\theta}_{n,k} - \theta_k\right), \qquad (6)$$

$$\mathrm{RMSE}(n) = \sqrt{\frac{1}{R}\sum_{k=1}^{R}\left(\hat{\theta}_{n,k} - \theta_k\right)^2}. \qquad (7)$$

In these equations, θ_k is the true score of the kth respondent, R is the total number of respondents, and $\hat{\theta}_{n,k}$ is the estimated ability value of the kth respondent, after applying n items. These values were captured in the selection of the first 30 items and after executing each CAT.

Setup and evaluation of the ALICAT approach

After executing the ISR evaluation, it is possible to know which is the rule with the best general performance to estimate user scores. The same is true for the ones with the best performance for users with high and low ability level. Ability values above or equal to 2.5 are considered high, and values below or equal to − 1.5 are considered low. After defining the above rules, it is possible to build the ALICAT.

The next step is to evaluate the ALICAT performance and compare it with other CATs that use only one ISR. This step follows the same process defined for evaluating the ISRs. After identifying the length of ALICAT, the tests will be processed again for all respondents in the sample, but with a fixed number of questions. We compute the values of the BIAS and the RMSE during the test execution.

Finally, as part of the performance comparison between ALICAT and CATs that exclusively uses one item selection rule, we will present graphs as the one in Fig. 3. This figure contains an example of a comparison between the true scores and the scores obtained via CAT. It can be seen in this example that the scores obtained by the CAT had the same values as the true scores, which would be the ideal case.

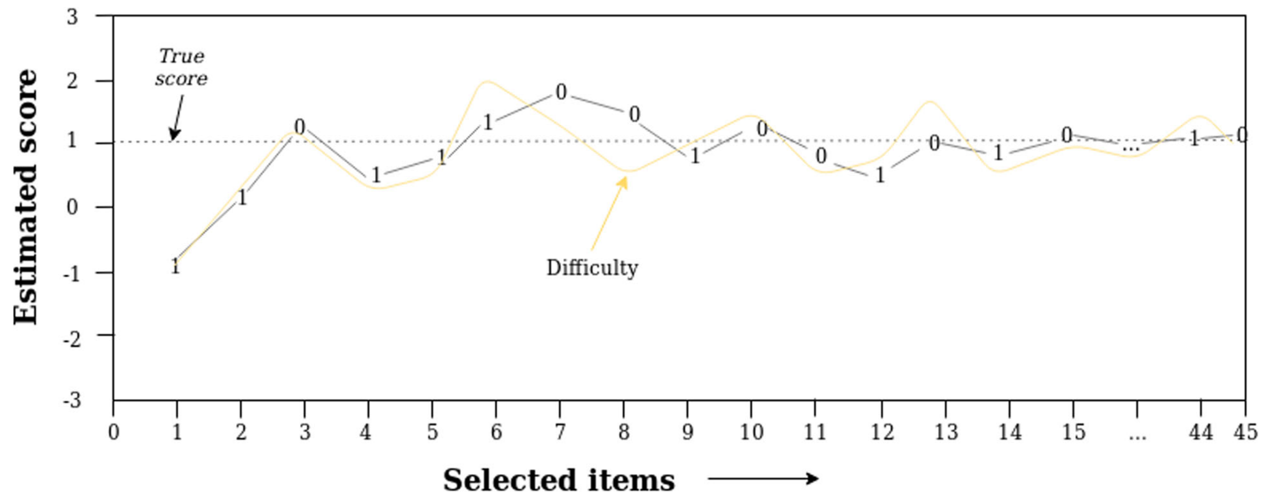


Fig. 2 Example of a CAT with 45 items applied to a respondent with $\theta = 1.1$

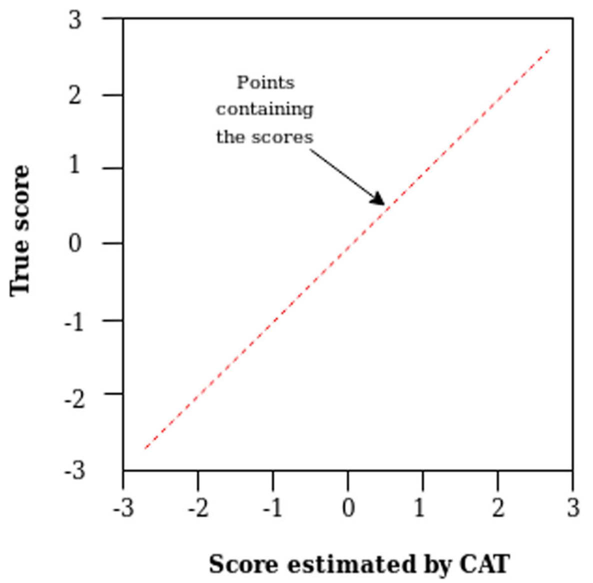


Fig. 3 Ideal scenario. This graph illustrates what would be an ideal scenario for estimating the respondents' abilities performed by a CAT. The red dashed line represents possible score values obtained by the CAT in relation to the true score

Results

The results were extracted from the execution of each CAT, considering its respective ISR. We considered their general performance, as well as their respective stability points. Also, the BIAS and RMSE were calculated in the initial stages and at the end of each CAT. Finally, the results of the ALICAT setup and execution stages are presented.

ISR performance in complete test

Table 2 contains the number of respondents (σ) and the average number of selected items ($\bar{x}$) for each $\hat{\theta}$ interval.

Table 2 Stability point identification (average number of selected itens $\bar{x}$) and sample size (σ) for each ISR and for each $\hat{\theta}$ interval

Estimated $\hat{\theta}$ interval	F		KL		KLP		MLWI		MPWI	
	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$
[− 2 ; − 1.5]	502	7	48	14	301	8	0	–	0	–
[− 1.5 ; − 1]	439	8	548	8	552	8	3	**23**	797	8
[− 1 ; − 0.5]	611	10	477	8	740	9	7	17	723	8
[− 0.5 ; 0]	490	14	1047	7	385	13	2544	6	517	9
[0 ; 0.5]	666	21	629	8	506	14	771	6	585	15
[0.5 ; 1]	281	19	369	10	708	9	377	7	395	13
[1 ; 1.5]	665	**35**	406	**21**	242	**24**	166	10	454	**18**
[1.5 ; 2]	152	25	554	12	1512	7	35	19	1506	9
[2 ; 2.5]	1142	19	901	15	3	22	769	14	2	2
[2.5 ; 3]	31	27	0	–	0	–	307	9	0	–

The values in bold were the maximum lengths defined for each ISR (denoted by n). With this new value, all ISRs were re-executed considering the stop criterion set to n.

In general, the *KLP* and *MPWI* rules practically failed to estimate the ability of users with high-level values ($\theta \geq 2$). With this, they placed all respondents with true θ greater than 1.5 in the range [1.5; 2]. At the other end, the *MLWI* rule had little success in estimating the ability of users with low-level values ($\theta \leq -0.5$).

The selection rule *F* was the one that selected, on average, a larger number of items for a given $\hat{\theta}$ group, and the *MPWI* was the one that selected the least ($n = 35$ and $n = 18$ items, respectively). Practically all the maximum averages of the number of selected items were found in the range [1; 1.5], except for the *MLWI* selection rule. The result obtained for rule *F* is similar to that found in [37], in which the maximum average of items was 33.

Figure 4 shows the performance of each item selection rule related to the number of questions selected for each $\hat{\theta}$ group. In most cases, rule *F* has the highest average number of selected items for $\hat{\theta}$ greater than -0.5. In contrast, it presents the best performance with users with $\theta \leq -1.5$. The *MLWI* rule is the best for $\hat{\theta}$ values between -0.5 and 1.5 and values greater than 2.5.

Table 3 shows the results of the BIAS and RMSE measures. There is evidence that the *KL* and *MPWI* rules are underestimating the ability of the respondents, since they have negative BIAS. The rules with the lowest RMSE, i.e., the ones with the best estimators, are *F* and *KLP*.

In general, the most prominent selection rule is *KLP*, since it has the lowest BIAS, the second lowest RMSE, and it allows to reduce the size of the test by 46.6%, with no significant loss of the respondents' estimated score, compared to the complete test with 45 questions.

Performance of the ISRs at the CATs' early stage

Next, we show the value of the BIAS and RMSE measures in the selection of the first 30 items during the CAT execution.

Figure 5 summarizes the results of the BIAS (Eq. 6) computed at the initial stage of the test for each ISR. Generally, the performance difference among the rules gets smaller as the number of questions increases. At the 30th question, for example, all rules have very similar performance.

The largest difference in BIAS values occurs in the most extreme groups of θs. For the negative extreme ($-2 \leq \theta < -1$), the rule *F* has a lower BIAS in the first 10 items. In the positive extreme ($2 < \theta \leq 3.5$), there was substantial variation. Although the *KL* rule performed well on items 2 and 3, we highlight the performance of the *MLWI* rule in that group. For instance, this rule presented a near zero result from items 4 to 10.

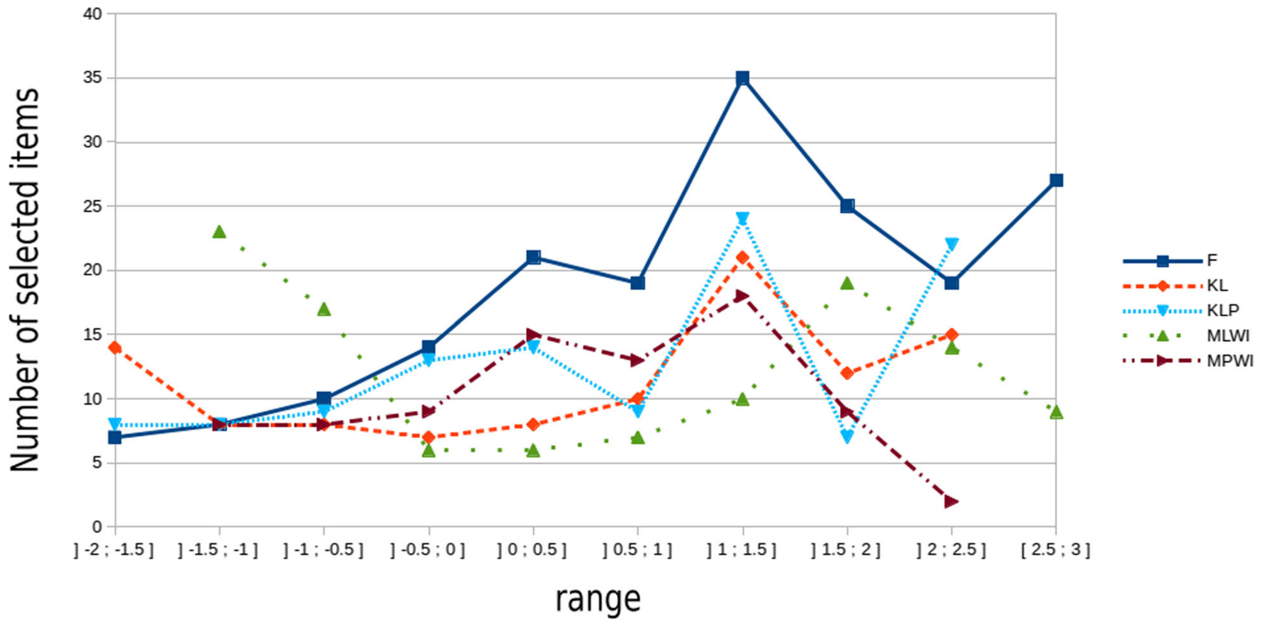


Fig. 4 Average number of selected items by ISRs for each estimated score interval

With the application of more items, we can observe that there is a BIAS convergence pattern that remains similar to the results presented in [12]. But the results were divergent especially for the *F* rule in negative extremes. This can be explained by the difference in the nature of the tests, by how the CAT was configured and also by the choice and setup of the abilities' estimation method.

Figure 6 shows that the RMSE behavior of the different ISRs, for the extreme θ values, was similar to that observed for the BIAS. The *F* and *MLWI* rules obtained a better RMSE for the extreme negative and positive θ values, respectively. However, the results exhibit the same divergent features when compared to the results shown in [12], which were already explained for the BIAS measure.

Configuration and performance analysis of the ALICAT approach

With the results of the last two subsections, it was possible to build the ALICAT. This approach is under the following definitions: The ISR with the general best performance is *KLP*; the ISR with the best performance for users with high-level abilities ($\theta \geq 2.5$) is *MLWI*; and the ISR with the best performance for users with low-level abilities ($\theta \leq -1.5$) is *F*. Thus, the ALICAT configuration modeled in Fig. 1 can be seen as its final version in Fig. 7 for this case study.

With the final design of the ALICAT completed, we executed it considering the complete test, the sample of the 4979 respondents, $p = 3$ and $q = 10$. The stability point identified was **21** questions. The results of Table 4 show that all $\hat{\theta}$ ranges have been contemplated. In addition, the number of respondents per group is very close to the 500 originally taken from the true θ values.

With the defined stability point, the ALICAT was then executed again, this time considering the fixed value of 21 questions as a stopping criterion. In this execution, the value of the BIAS obtained was **0.004** and the RMSE was **0.190**. These values are very close to the results of *KLP*, which obtained the best general performance.

Figure 8 displays the results of CAT scores versus true scores considering a complete test. The ALICAT and the CAT with exclusive use of the *KLP* rule had similar results. The CATs with the *KL* and *MLWI* rules had many divergences in the estimations of scores below 0 (zero). The biggest difference in the *MPWI* estimations was for θ values lesser than 0 and greater than 2. The main difference between the ALICAT and the *F* rule was for θ values closer to 3. While *F* had an apparent best performance, it was also the one that required more items to converge. Altogether, there were required 35 items, compared to 21 used in the ALICAT execution.

Table 3 BIAS and RMSE for each item selection rule

	F	KL	KLP	MLWI	MPWI
BIAS	0.030	− 0.002	0.001	0.067	− 0.028
RMSE	0.174	0.273	0.193	0.400	0.294

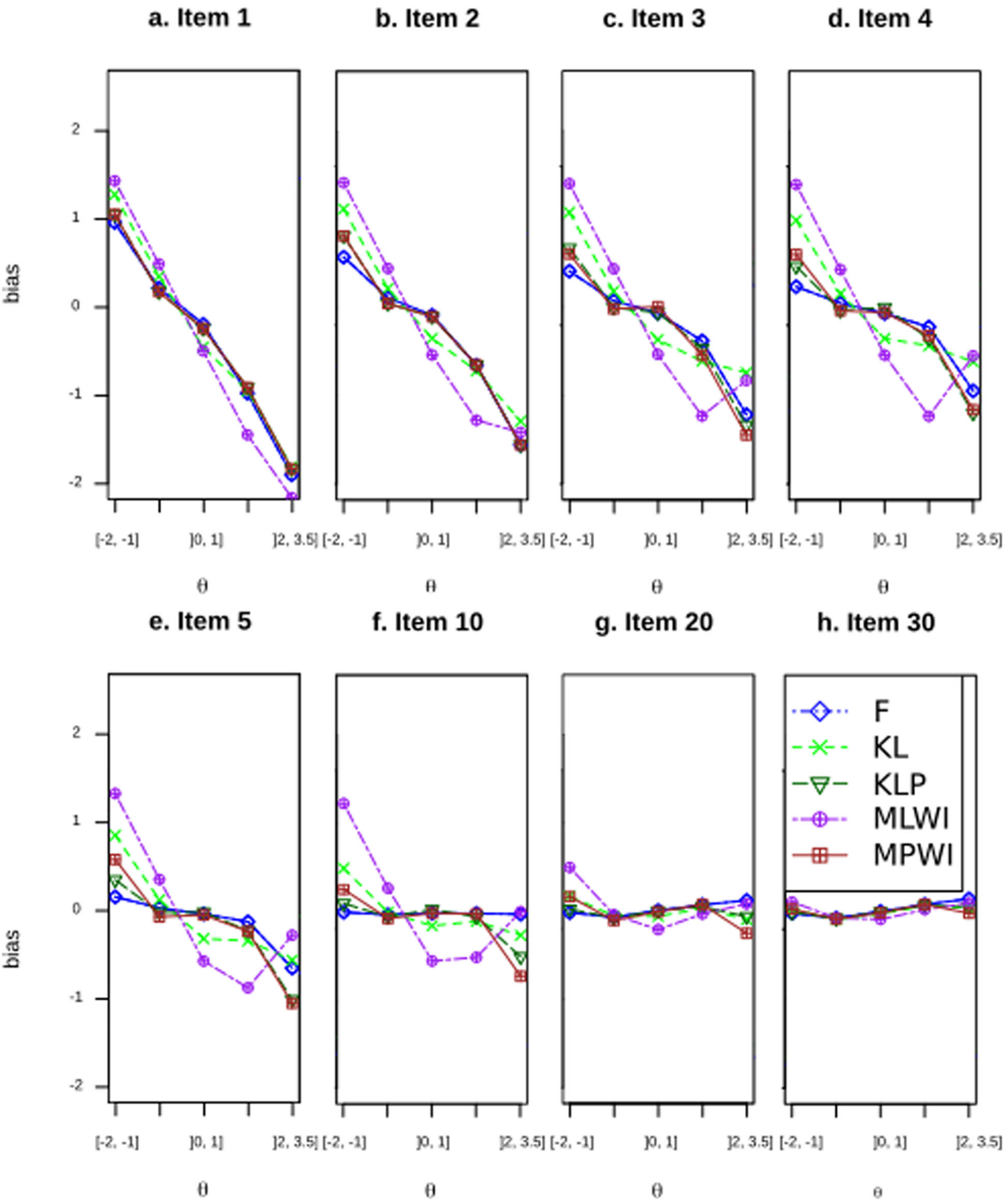


Fig. 5 BIAS obtained from the execution of the CATs for each ISR. The BIAS were cataloged by groups of θ for the first 30 items. Groups range from − 2 to 3.5 (*x*-axis)

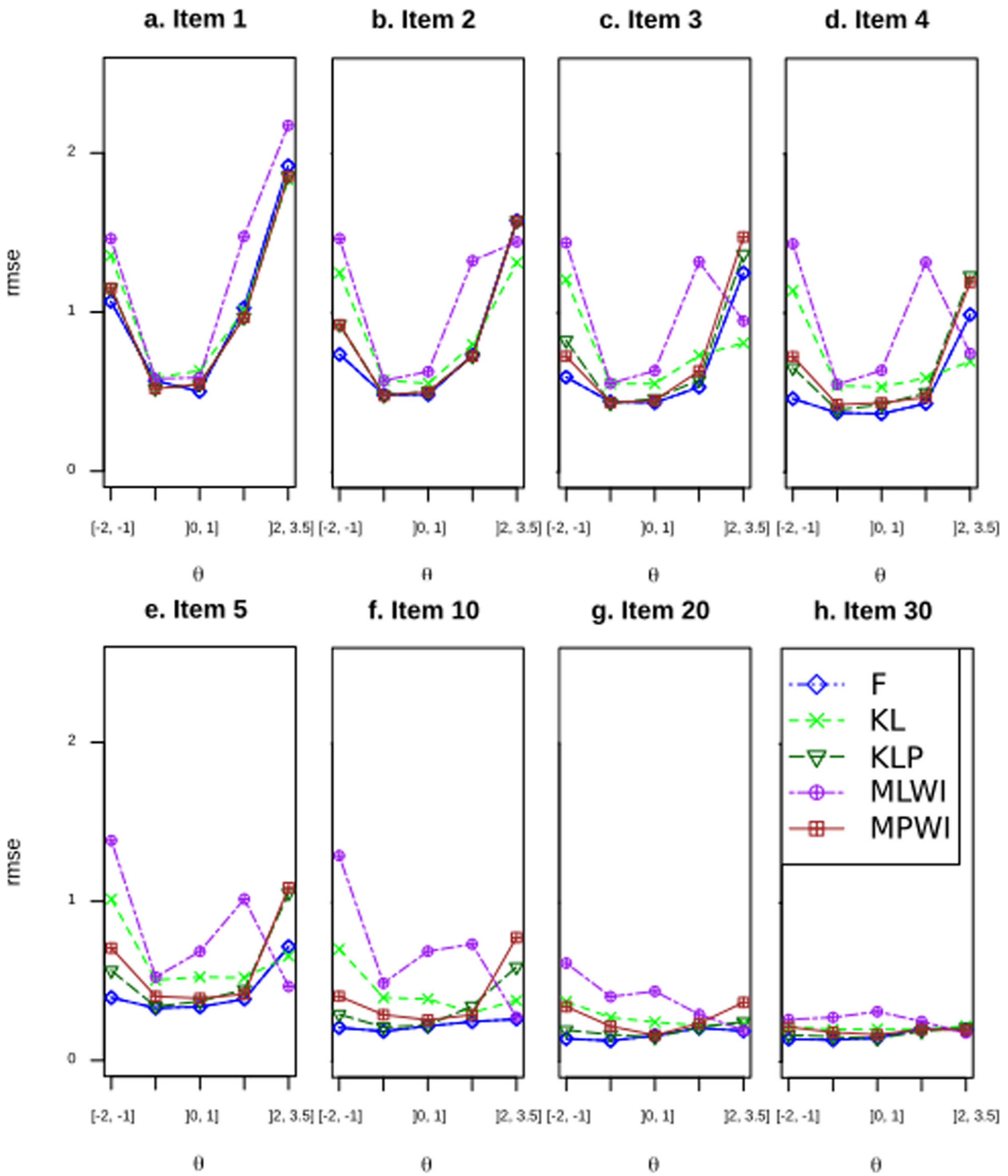


Fig. 6 RMSE obtained from the execution of the CATs for each ISR. The RMSE were cataloged by groups of θ for the first 30 items. Groups range from − 2 to 3.5 (*x*-axis)

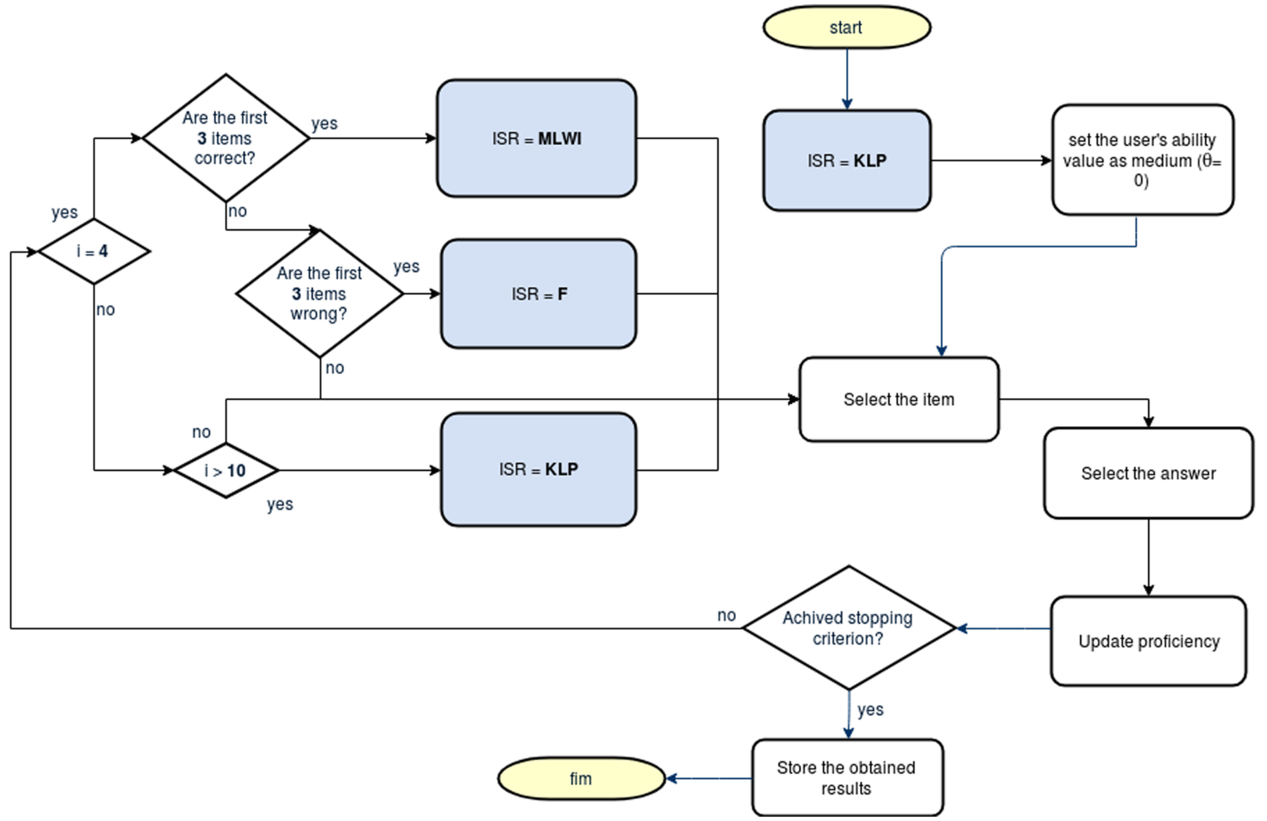

start
ISR = **KLP**
set the user's ability value as medium (θ= 0)
Are the first **3** items correct?
yes
ISR = **MLWI**
no
yes
i = **4**
no
Are the first **3** items wrong?
yes
ISR = **F**
no
Select the item
Select the answer
no
i > **10**
yes
ISR = **KLP**
Update proficiency
no
Achived stopping criterion?
yes
Store the obtained results
fim

Fig. 7 Flowchart of the final ALICAT configuration after the analysis of the ISRs

Conclusions and future works

The *F* and *KLP* rules have good general performances in the estimation of θ values considering the point of stability as the stopping criterion. However, *KLP* required fewer questions (24 items versus 35 that rule *F* required) making it the best option in the overall ISR evaluation. When investigating the behavior of the ISRs in the initial moments of the CATs, *MLWI* and *F* had the best performances for positive and negative extreme θ values, respectively. Based on these results, it was possible to configure and validate the ALICAT method. Its performance for BIAS and RMSE was close to *KLP*; however, it required only 21 items. This represents a 54.3% reduction over the 45 items in the *Math and its Technologies* test of the 2012 ENEM exam with no significant loss in ability estimation.

In summary, the ALICAT was able to further optimize the item selection process by dynamically choosing the ISR, rather than the fixed strategy of using only one ISR throughout the test. This improvement in the process of constructing the computerized adaptive tests represents a direct reduction in the test's resolution time. This feature covers advantages, such as (i) cost reduction for the institution that is applying the test, because users will spend less time using the computational resources and the physical spaces, if these are applied; and (ii) decreased fatigue and frustration of respondents. These characteristics can improve the motivation of the respondents in the resolution of the items, and thus have a better CAT accuracy to estimate their abilities. This is possible because, among other aspects, participants will not have to respond to very easy or very difficult questions related to their level of knowledge.

Table 4 Number of respondents (σ) and average selected items ($\bar{x}$) for each $\hat{\theta}$ interval in ALICAT's approach

$\hat{\theta}$ Interval	ALICAT	
	σ	$\bar{x}$
[− 2 ; − 1.5]	495	7
[− 1.5 ; − 1]	471	8
[− 1 ; − 0.5]	626	10
[− 0.5 ; 0]	540	11
[0 ; 0.5]	507	13
[0.5 ; 1]	754	9
[1 ; 1.5]	449	18
[1.5 ; 2]	100	18
[2 ; 2.5]	554	**21**
[2.5 ; 3]	483	10

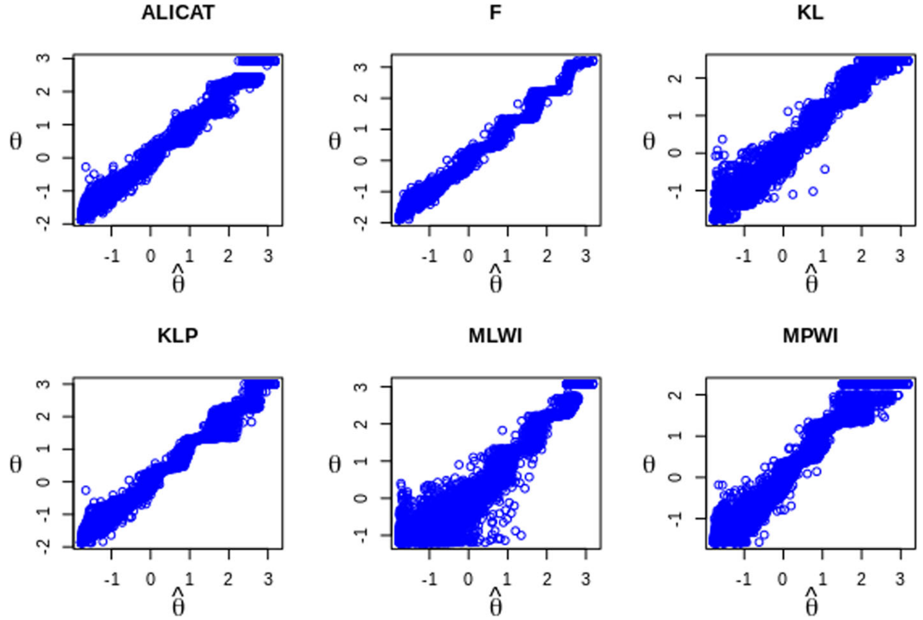


Fig. 8 Comparison between CAT scores ($\hat{\theta}$) and true scores (θ)

In the context of this case study, we used only data from one macro area of the ENEM exam. However, this study may be extended to other areas of knowledge or to other tests of a similar nature.

One limitation of this work is the form used to define the maximum length n of the test. We defined it as the highest average value of the stability points for the 10 groups of $\hat{\theta}$ values. This may decrease the accuracy in estimating respondents who have a higher point of stability.

A preliminary version of this paper, which does not include anything about the ALICAT and the performance evaluation of the ISRs at the CATs' early stage, was published in [40].

For future works, we want to apply the ALICAT in different educational tests and use different methods for defining the CAT length and its configuration. In addition, other methods of item selection can also be verified in the comparative study of ISRs. Thus, it will allow to validate the effect on the score estimation of the ALICAT approach in different test scenarios.

Abbreviations

CAT: Computerized adaptive testing; ALICAT: Personalized computerized adaptive testing; IRT: Item response theory; ENEM: *Exame Nacional do Ensino Médio*, National High School Exam; P&P: Paper and pencil; ISR: Item selection rule; CTT: Classical test theory; BI: Bank of items; ML3: Logistic model with three parameters; F: Fisher information; KL: Kullback-Leibler information; KLP: Kullback-Leibler information with a posterior distribution; MLWI: Maximum likelihood weighted information; yMPWI: Maximum posterior weighted information; INEP: *Instituto Nacional de Educação e Pesquisas Educacionais Anísio Teixeira*, National Institute for Educational Studies and Research Anísio Teixeira; EAP: Expected a posteriori; SE: Standard error; RMSE: Root mean squared error

Acknowledgements

The authors would like to thank the funding agency.

Authors' contributions

VMGJ, JSF, and VF made substantial contributions to the conception and design of the ALICAT approach. They also were responsible for the acquisition of data, analysis, and interpretation. ASR was a major contributor in writing the manuscript and reviewing the data analysis. KVD also contributed to the conception and design and has also been involved in revising it critically for important intellectual content. All authors read and approved the final manuscript.

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Availability of data and materials

As already mentioned in our "Research method" section, the data sets analyzed during the current study are from the 2012 ENEM exam, which are public and have been taken from the transparency portal [36]. All the remaining data generated during this study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Information Systems, School of Arts, Sciences and Humanities, University of São Paulo, Rua Arlindo Bettio n^{o}1000, São Paulo 03828-000 Brazil. [2] Exact Science and Earth Department, University of Bahia State (UNEB), Rua Silveira Martins, 2555, Cabula, Salvador, 41150-000, Brazil.

Received: 4 January 2019 Accepted: 1 May 2020
Published online: 19 May 2020

References

1. Guzmán E, Conejo R (2004) A model for student knowledge diagnosis through adaptive testing. In: International Conference on Intelligent Tutoring Systems. Springer, Berlin. pp 12–21
2. Kovatcheva E, Nikolov R (2009) An adaptive feedback approach for e-learning systems. IEEE Technol Eng Educ(ITEE) 4(1):55–57
3. López-Cuadrado J, Pérez TA, Vadillo J. Á., Gutiérrez J (2010) Calibration of an item bank for the assessment of Basque language knowledge. Comput Educ 55(3):1044–1055

4. Wong K, Leung K, Kwan R, Tsang P (2010) E-learning: developing a simple web-based intelligent tutoring system using cognitive diagnostic assessment and adaptive testing technology. In: Tsang P, Cheung SKS, Lee VSK, Huang R (eds). Hybrid learning. Springer, Berlin, Heidelberg. pp 23–34
5. de Andrade DF, Tavares HR, da Cunha Valle R (2000) Teoria da resposta ao item: conceitos e aplicações. ABE - Associação Brasileira de Estatística, São Paulo
6. Instituto Nacional de Educação e Pesquisas Educacionais Anísio Teixeira Entenda sua nota no ENEM (2013). http://download.inep.gov.br/educacao_basica/enem/guia_participante/2013/guia_do_participante_notas.pdf. Accessed: 29 Mar 2017
7. Chen C-M, Lee H-M, Chen Y-H (2005) Personalized e-learning system using item response theory. Comput Educ 44(3):237–255
8. Segall DO (2004) A sharing item response theory model for computerized adaptive testing. J Educ Behav Stat 29(4):439–460
9. Wainer H (2000) Computerized adaptive testing. Routledge, New York, NY
10. Bernes-Lee T (2006) Linked data - design issues. http://www.w3.org/DesignIssues/LinkedData.html. Accessed 4 Dec 2018
11. Chen S-Y, Ankenman RD (2004) Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. J Educ Meas 41(2):149–174
12. Chen S-Y, Ankenmann RD, Chang H-H (2000) A comparison of item selection rules at the early stages of computerized adaptive testing. Appl Psychol Meas 24(3):241–255
13. Barrada JR, Olea J, Ponsoda V, Abad FJ (2008) Incorporating randomness in the Fisher information for improving item-exposure control in cats. Br J Math Stat Psychol 61(2):493–513
14. Wang C, Chang H-H, Huebner A (2011) Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. J Educ Meas 48(3):255–273
15. Olea Diaz J, Ponsoda V (2013) Tests adaptativos informatizados Editora UNED AULA. ABIERTA, Spain
16. Spenassato D, Bornia AC, Tezza R (2015) Computerized adaptive testing: a review of research and technical characteristics. IEEE Lat Am Trans 13(12):3890–3898
17. Thompson NA, Weiss DJ (2011) A framework for the development of computerized adaptive tests. Pract Assess Res Eval 16(1):1–9
18. Al-A'Ali M (2007) Implementation of an improved adaptive testing theory. Educ Technol Soc 10(4):80–94
19. Reckase MD, Weiss D (1983) A procedure for decision making using tailored testing. In: New horizons in testing: latent trait test theory and computerized adaptive testing. Academic Press, San Diego. pp 237–255
20. Collins JA, Greer JE, Huang SX (1996) Adaptive assessment using granularity hierarchies and Bayesian nets. In: International conference on intelligent tutoring systems. Springer, Berlin. pp 569–577
21. Rudner LM (2002) An examination of decision-theory adaptive testing procedures. In: Annual Meeting of the American Educational Research Association, New Orleans. pp 1–14
22. Hambleton RK, Swaminathan H, Rogers HJ (1991) Fundamentals of item response theory. Sage, Newbury Park, California
23. Galvao AF, Neto RF, Borges CCH (2013) Um modelo inteligente para seleção de itens em testes adaptativos computadorizados. Master's thesis. Universidade Federal de Juiz de Fora (UFJF), Brazil
24. Eggen T, Straetmans G (2000) Computerized adaptive testing for classifying examinees into three categories. Educ Psychol Meas 60(5):713–734
25. Van Rijn P, Eggen T, Hemker B, Sanders P (2002) Evaluation of selection procedures for computerized adaptive testing with polytomous items. Appl Psychol Meas 26(4):393–411
26. Butterfield MS (2016) Comparing item selection methods in computerized adaptive testing using the rating scale model. PhD thesis
27. Lord FM (1980) Applications of item response theory to practical testing problems. Routledge, New York, NY
28. Chang H-H, Ying Z (1996) A global information approach to computerized adaptive testing. Appl Psychol Meas 20(3):213–229
29. der Linden WJV, Pashley PJ (2009) Item selection and ability estimation in adaptive testing. In: Elements of adaptive testing. Springer, New York, NY. pp 3–30
30. van der Linden WJ (1998) Bayesian item selection criteria for adaptive testing. Psychometrika 63(2):201–216
31. Veerkamp WJ, Berger MP (1997) Some new item selection criteria for adaptive testing. J Educ Behav Stat 22(2):203–226
32. Bock RD, Mislevy RJ (1982) Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas 6(4):431–444. https://doi.org/10.1177/014662168200600405
33. Gonçalves JP, Aluísio SM (2015) Teste adaptativo computadorizado multidimensional com propósitos educacionais: Princípios e métodos. Rev Ensaio: Avaliação e Políticas Públicas em Educação 23(87):389–414
34. Costa CES (2015) Análise da dimensionalidade e modelagem multidimensional pela tri no enem (1998-2008). PhD thesis, Universidade Federal de Santa Catarina
35. Andrade GG (2012) A metodologia do enem: uma reflexão. Série-Estudos-Periódico do Programa de Pós-Graduação em Educação da UCDB 33:67–76
36. (2016) Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira: Microdados do ENEM. Brasília: Inep, 2016. http://portal.inep.gov.br/basica-levantamentos-acessar. Accessed: 29 Mar 2017
37. Spenassato D, Trierweiller AC, de Andrade DF, Bornia AC (2016) Testes adaptativos computadorizados aplicados em avaliações educacionais. Rev Bras de Informatica na Educação 24(2):1–12
38. Hanson BA (2002) IRT command language. http://www.openirt.com/b-a-h/software/irt/icl/. Accessed 4 Dec 2018
39. Magis D, Raîche G, et al (2012) Random generation of response patterns under computerized adaptive testing with the R package catR. J Stat Softw 48(8):1–31
40. Jatobá V, Delgado KV, Farias J, Freire V (2018) Comparação de regras de seleção de itens em testes adaptativos computadorizados: um estudo de caso no enem. In: Proceedings of the Brazilian Symposium on Computers in Education, Fortaleza. pp 1453–1462

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.