Publication status: Not informed by the submitting author

Progressive-restricted method decreases item exposure of a Computerized Adaptive Testing without compromising precision

Alexandre Jaloto, Ricardo Primi

https://doi.org/10.1590/SciELOPreprints.6578

Submitted on: 2023-08-07
Posted on: 2023-08-28 (version 2)
(YYYY-MM-DD)

Progressive-restricted method decreases item exposure of a Computerized Adaptive Testing without compromising precision

Alexandre Jaloto (Inep, Brazil)

https://orcid.org/0000-0002-5291-1768

Ricardo Primi (USF, Brazil)

https://orcid.org/0000-0003-4227-6745

Abstract

The Progressive-Restricted method (PR) increases the security (in terms of item exposure) of Computerized Adaptive Testing (CAT). However, little is known about the effect of the acceleration parameter on precision and security. Therefore, we evaluated the PR with different acceleration parameters in CATs with fixed and variable-length. We combined item selection methods (Maximum Fisher Information – MFI – and PR) with stopping criteria (Fixed length, Standard error of 0.30 and Error reduction of 0.015) and simulated CATs for each Enem test. With the error reduction criterion, the precision with MFI was lower. In the other CATs, the precision was similar. Security has increased with larger acceleration parameters. At last, we compared the linear format of the Enem with the 20-item CAT. The latter had greater precision.
Keywords: psychometrics; large-scale educational assessment; computerized adaptive test; simulation; item exposure control.

Computerized Adaptive Testing (CAT) can increase the efficiency of a test by reducing its length (Bulut & Kan, 2012; Mizumoto et al., 2019; Spenassato et al., 2016). However, in high-stakes tests, test security (in terms of item exposure) must also be considered, along with efficiency. Through a random component, the progressive-restricted (PR) method controls the exposure of items with little impact on efficiency and precision (Leroux & Dodd, 2016; Leroux et al., 2013). The importance of this random component in item selection decreases over the application, and the speed of this decrease impacts the amount of low-exposed items (Barrada et al., 2008). This speed is determined by the acceleration parameter in the PR method equation, which has been little explored, especially in variable-length CAT. Therefore, in this study, we evaluated the use of PR exposure control with different acceleration parameters in fixed-length

and variable-length CATs using data from the Brazilian National High School Exam (Enem), a high-stakes educational test.

CAT can optimize test administration because items are administered to the participant based on their response to the previous item (Weiss & Kingsbury, 1984). When the Maximum Fisher Information (MFI) method is used, the selected item is the one with the highest information for the provisional theta (latent variable) of the examinee. Therefore, CAT increases the efficiency of test administration (in terms of test length) without compromising measurement precision, or even improving it. For example, the study by Sulak and Kelecioğlu (2019) had an average test length of 7.07 items with a bank of 250 items. Spenassato et al. (2016) simulated the administration of the Enem 2012 in CAT format and reduced the test length from 45 to 33 items, with a correlation of 0.998 between the original and simulated thetas. In high-stakes educational tests like the Enem (which is used for higher education admission in Brazil), in addition to efficiency and precision, test security is a factor that needs to be considered. One way to increase test security is to control item exposure.

The PR exposure control method avoids item overexposure and reduces item underexposure and overlap. Overlap corresponds to the proportion of identical items administered to two randomly selected examinees. In the simulation by Leroux and Dodd (2016), conditions without exposure control had the root mean square error (RMSE) reaching 0.31. When the PR method was implemented, this value did not exceed 0.34. In addition to the small decrease in precision, this method provided greater test security, as the lowest overlap rate obtained without exposure control was around 0.42, while with PR the highest rate was around 0.21. In other words, without using PR, two random examinees shared about 42% of their test items. This rate was halved when this method was adopted. Leroux et al. (2013) obtained similar results in simulations of both fixed-length and variable-length CATs. The RMSE value increased by a maximum of 0.04, and the overlap rate decreased from 0.54 to 0.25

when PR was implemented. Other studies have also pointed to the potential of using PR to increase test security without compromising precision compared to the MIF method (e.g., Lee & Dodd, 2012; Leroux et al., 2019).

The PR combines two methods of exposure control: restricted maximum information and progressive (Revuelta & Ponsoda, 1998). In the restricted maximum information method, a maximum exposure rate is established for items. At the beginning of the administration, only items with exposure rates lower than the maximum are available for administration. The remaining items are selected using the MFI method.

In the progressive method, the administered item is the one that has the highest sum between two components: a random component and an informative component. At the beginning of the administration, the importance of the random component is 100% and that of the informative component is 0%. The importance of the random component decreases over the application, while that of the informative component increases. In terms of notation, in the progressive method, the selected item j^* will be the one that:

$$j^* = arg\max_{j \in S} \left| (1 - W)R_j + WI_j\left(\hat{\theta}_t\right) \right| \quad (1)$$

where S represents the bank of items available for administration, $\hat{\theta}_t$ corresponds to the provisional theta after the administration of t items, $I_j\left(\hat{\theta}_t\right)$ is the information of item j for the provisional theta, R_j is a random number drawn from a uniform distribution with a range between zero and the value of the highest information among the available items in the bank, $\left[0; max_{j \in S} I_j\left(\hat{\theta}_t\right)\right]$, and W é is the weight that determines the importance of the random and informative components in the equation for item j. The expression $arg\max_{j \in S}$ indicates that item j with the highest result of the function will be selected, i.e., the one that presents the highest sum between the random and informative components.

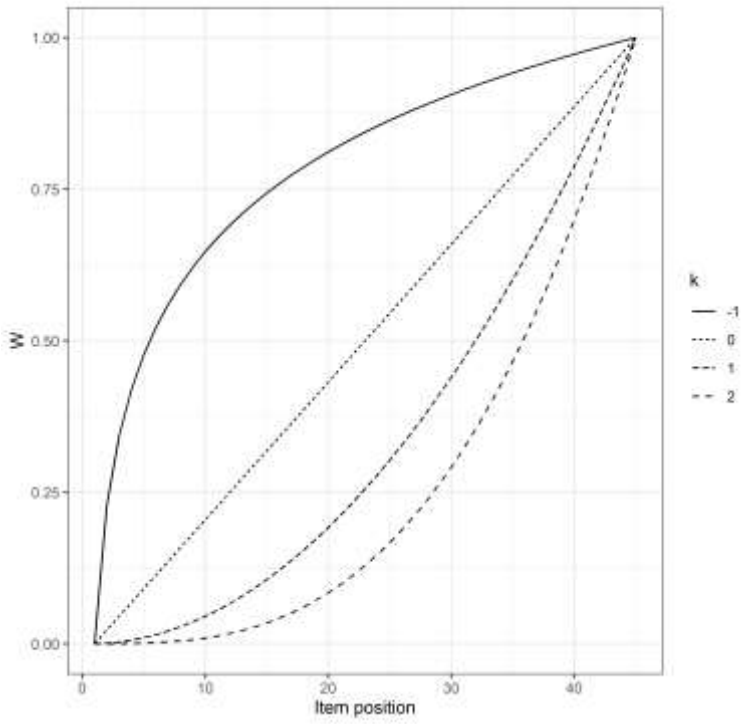
The weight W can be calculated as follows (Barrada et al., 2008):

$$W = \begin{cases} 0, & \text{if } t = 0 \\ \frac{\sum_{b=2}^{t}(b-1)^k}{\sum_{b=2}^{N}(b-1)^k}, & \text{if } t \neq 0 \end{cases} \quad (2)$$

where N is the test length and k is the accelerator parameter. This parameter affects the speediness of increase of W, that is, the rate at which W moves away from 0 during the administration. The larger the parameter k, the slower W increases, and the slower the random component loses importance. Figure 1 illustrates four situations of k values in a 45-item CAT.

Figure 1

Variation of the weight in the progressive-restricted method with different acceleration parameters, in fixed-length CATs



For variable-length tests, W can be calculated as follows (Magis & Barrada, 2017):

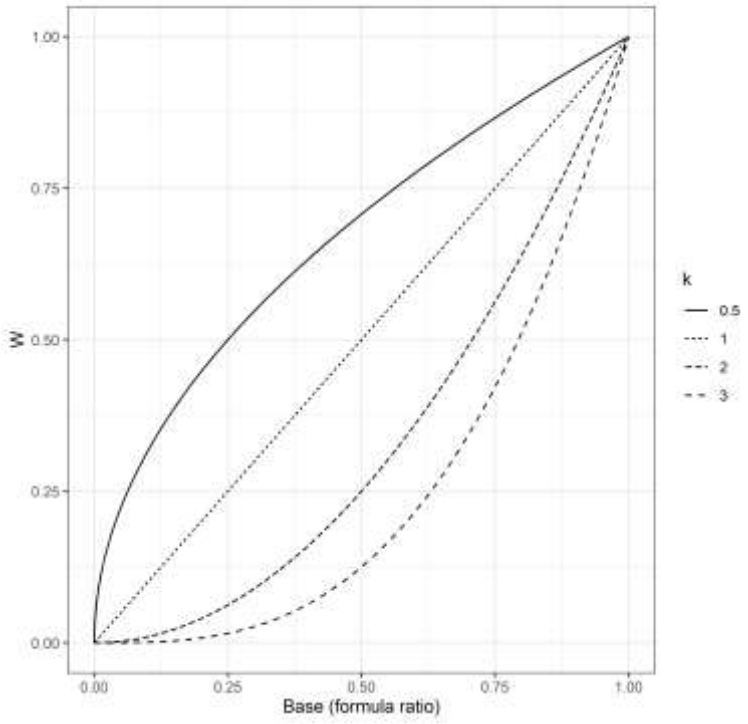
$$W = \begin{cases} 0, & \text{if } t = 0 \\ max\left[\frac{I(\hat{\theta}_t)}{I_{stop}}, \frac{q}{M-1}\right]^k, & \text{if } t \neq 0 \end{cases} \quad (3)$$

where I_{stop} is the information needed to reach the stopping value of the standard error and M is the maximum test length.

There are two situations in Equation 3 that affect the weight of the random component of Equation 1 undesirably: when $k = 0$ and when $k < 0$. When $k = 0$, from the second item onward, $W = 1$ constantly (Equation 3). In this case, the weight of the random component will always be zero starting from the selection of the second item (Equation 1), and items will be selected solely based on their psychometric information. On the other hand, when $k < 0$, the value of W decreases as the administration progresses (Equation 3), and consequently, the weight of the random component increases (Equation 1). The effect is that the weight of the item information becomes smaller as the administration progresses. Therefore, in this work, we only used positive values of k. Figure 2 illustrates four situations of k values in a variable-length CAT. The x-axis corresponds to the formula ratio, which is the base in Equation 3, i.e., $max\left[\frac{I(\hat{\theta}_t)}{I_{parada}}, \frac{q}{M-1}\right]$.

Figure 2

Variation of the weight in the progressive-restricted method with different acceleration parameters, in variable-length CATs



Note. x-axis corresponds to the base in Equation 3.

From Figure 1 and Figure 2, it can be observed that as the acceleration parameter k increases, W increases more slowly, and the random component loses importance more slowly. At the same time, the importance of the information increases more slowly. Consequently, the importance of item discrimination (which is directly related to its information) will also increase more slowly. Therefore, less discriminative items are more exposed under these conditions compared to situations where the importance of information is higher (Barrada et al., 2008). Increasing the exposure of less discriminative items may reduce the demand for new items in the item bank, as exposure of other items will decrease. Although the change in the acceleration parameter may have a positive impact on the characteristics of a CAT, its effect on test efficiency, precision, and security has been little studied (e.g., Barrada et al., 2008). Additionally, the effect of this changing in variable-length CAT is not known.

Examples of stopping rules for a variable-length CAT are the standard error and the reduction of error after item administration. The combination of these criteria increases the efficiency of the CAT compared to using the standard error alone, as the application terminates when the administration of new items does not contribute to the reduction of measurement error. Kallen et al. (2018) reduced a CAT that originally had a stopping rule of standard error (0.3) and a maximum of 12 items. The authors included a reduction of error (0.01) as a stopping rule and found that the proportions of applications terminated by this criterion reached 13.3%, with little impact on precision. The length of the tests reduced by up to 3.3 items on average. Using a difference of 0.015 as the predicted error reduction value as a stopping rule also increases the efficiency of the CAT with little impact on precision when the item bank has a nonuniform distribution (Morris et al., 2020). The combination of the standard error criterion with the error reduction criterion increases the efficiency of the CAT, although it is little explored.

Given the gaps identified in the paragraphs above, this study selected the PR method to investigate to what extent it alters the efficiency (in terms of test length), precision, and security (in terms of item exposure) of a CAT with different stopping rules. Our aim was to evaluate the PR exposure control with different acceleration parameters in fixed-length and variable-length CATs. We simulated the administration of the Brazilian Enem (a high-stakes educational test administered on paper) in a CAT format. The Enem, which is administered annually by the National Institute for Educational Studies and Researches Anísio Teixeira (Inep), is composed of four tests with 45 multiple-choice items each, namely: Human Sciences and its Technologies (HS), Natural Sciences and its Technologies (NS), Languages, Codes, and its Technologies (LC), and Mathematics and its Technologies (MT).

Our research questions were as follows:

1. How do the efficiency, precision, and security of the test vary when adopting the PR exposure control?
2. How do the efficiency, precision, and security vary as a function of the acceleration parameter of the PR method?
3. What is the impact on the precision of the Enem when reducing its length through a CAT with exposure control?

We formulated the following study hypotheses: (H1) the efficiency and precision of the PR method will be similar to those of MFI; (H2) the security of PR will be higher than that of MFI; (H3) the efficiency and precision of the PR method will be similar across all acceleration parameters; (H4) the higher the acceleration parameter, the higher the security; and (H5) the precision of a CAT with PR will be higher than that of a linear test.

This study advances because it evaluates the PR method and different values of its acceleration parameter, which are underexplored in the literature, especially in variable-length CAT. Moreover, it is novel because it evaluates the combination of the PR method with the

stopping criterion of observed error reduction, which, despite being efficient, is also underexplored. Lastly, the simulation study uses robust item banks (over 700 items) previously administered in a real-world setting.

Methods

Study design

For each type of CAT (fixed-length and variable- length), we manipulated two variables: item exposure control and stopping criterion. As item exposure control, we used the PR method with two acceleration parameters: for fixed-length CATs, PR with $k = 1$ (PR1) and PR with $k = 2$ (PR2); for variable-length CATs, PR with $k = 2$ (PR2) and PR with $k = 3$ (PR3). Therefore, we had three exposure control conditions for each type of CAT, one of them being the absence of control. The maximum item exposure rate for PR was set at 0.30.

As item selection methods, we adopted the MFI (without item exposure control, and used in conjunction with the PR method) and the random method. The latter was used only for reference purposes, to quantify the improvement of the other conditions. We did not use item exposure control with the random method. Thus, we had a total of five item selection methods (random, MFI, PR1, PR2, and PR3), with four methods for each type of CAT. We used four stopping rules: for fixed-length CATs, 45 items (FL45) and 20 items (FL20); for variable-length CATs, standard error of 0.30 (SE30) and a combination of standard error of 0.30 with error reduction of 0.015 (ER015).

As we had four item selection methods, four stopping rules, replicated each bank 20 times, and applied the design to the four testes of the Enem, this study had a total of $4 \times 4 \times 20 \times 4 = 1280$ simulations. Table 1 shows the conditions of the simulations for fixed-length CATs and Table 2 shows the conditions of the simulations for variable-length CATs. All commands are available at http://github.com/alexandrejaloto/PR_CAT.

Table 1

Conditions of the simulations for fixed-length CATs

Stopping rule	Selection method			
	Random	Maximum Fisher information	PR with a 0.30 rate and AP 1	PR with a 0.30 rate and AP 2
Fixed length of 45 items	RANFL45	MFIFL45	PR1FL45	PR2FL45
Fixed length of 20 items	RANFL20	MFIFL20	PR1FL20	PR2FL20

Note. PR = progressive-restricted; AP = acceleration parameter of progressive method.

Table 2

Conditions of the simulations for variable-length CATs

Stopping rule	Selection method			
	Random	Maximum Fisher information	PR with a 0.30 rate and AP 2	PR with a 0.30 rate and AP 3
Standard error of 0.30	RANSE30	MFISE30	PR2SE30	PR3SE30
Standard error of 0.30 or Error reduction of 0.015	RANER015	MFIER015	PR2ER015	PR3ER015

Nota. PR = progressive-restricted; AP = acceleration parameter of progressive method.

Response bank

For each area of the Enem, we drew a simple random sample of participants from the 2020 edition. Data were obtained from the microdata of Enem (available at https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem) on 01/11/2022. The sample size was determined to ensure a mean with a standard error of 5 points on the Enem scale (equivalent to a coefficient of variation of 0.01) with a 95% confidence interval. The calculation of the sample size, denoted as n, was done as follows:

$$n = \left(\frac{\sigma \times Z}{d}\right)^2 \quad (4)$$

where σ is the population standard deviation, d is the acceptable error (5), and Z is the value of z in the normal distribution curve with an area of $[1 - 0{,}05/2]$. The value of Z that ensures a 95% confidence interval is approximately 1.96.

By adopting this procedure, we were able to generalize our results to the population of the 2020 edition of the Enem, which potentially brings our simulation closer to expected situations for future editions with similar characteristics. Additionally, we supplemented the

sample size until it was three times larger than the number of items in the item bank. This allowed for a more robust comparison of results related to item exposure, as it kept the ratio between the size of the item bank and the size of the participants constant. For example, if we had different proportions, the fact that some items were not presented in the simulation of an area could be due to the large size of the item bank relative to the sample size. As an illustration, in the LC area, the sample size needed to guarantee a standard error of 5 points is smaller than the size of the item bank.

The descriptive statistics of the participants in the Enem 2020 and the samples from each area are presented in Table 3. Although the official scores are reported in a metric with a mean of 500 and a standard deviation of 100 (with the reference being the participants of the Enem 2009), in this study, we used a standardized metric (with a mean of 0 and a standard deviation of 1) to facilitate the analyses.

Table 3

Descriptive statistics of the participants in the Enem 2020 and the simulation samples

Area	n	Mean (standard deviation)	Range
Human Sciences			
Participants in Enem	2,749,073	0.09 (0.83)	-1.67–3.22
Simulation sample	2,268	0.06 (0.82)	-1.53–2.43
Nature Sciences			
Participants in Enem	2,596,735	-0.09 (0.70)	-1.57–3.13
Simulation sample	2,247	-0.09 (0.71)	-1.48–2.51
Languages and Codes			
Participants in Enem	2,751,791	0.22 (0.68)	-1.95–2.79
Simulation sample	2,649	0.22 (0.67)	-1.79–2.07
Mathematics			
Participants in Enem	2,596,527	0.16 (0.90)	-1.33–3.66
Simulation sample	2,376	0.14 (0.90)	-1.33–3.35

The item responses were generated using Monte Carlo simulation through the gen.resp function of the simCAT package (Jaloto & Primi, 2023a). Each simulation resulted in a response matrix whose rows corresponded to the subjects in the sample, and whose columns corresponded to the items. In other words, we generated a response bank as if each subject had

answered all the items in an area. Each response bank was generated 20 times for the purpose of simulation replication.

CAT specifications

The item bank was composed of items administered in the Enem from 2009 to 2020, which were available in the microdata. The files were obtained on 11/01/2022, except for the microdata from Enem 2011, which were downloaded on 11/18/2022.

The content of the items is based on a list of 30 skill descriptors for each area. We excluded items that did not have information about the assessed content and those that were excluded by the Inep team from the analyses (these items did not have IRT parameters). The item banks ranged from 749 (NS) to 883 (LC) items. Table 4 presents the description of the four item banks. Figure 3 shows the test information curve and the density distribution of the thetas for each sample.

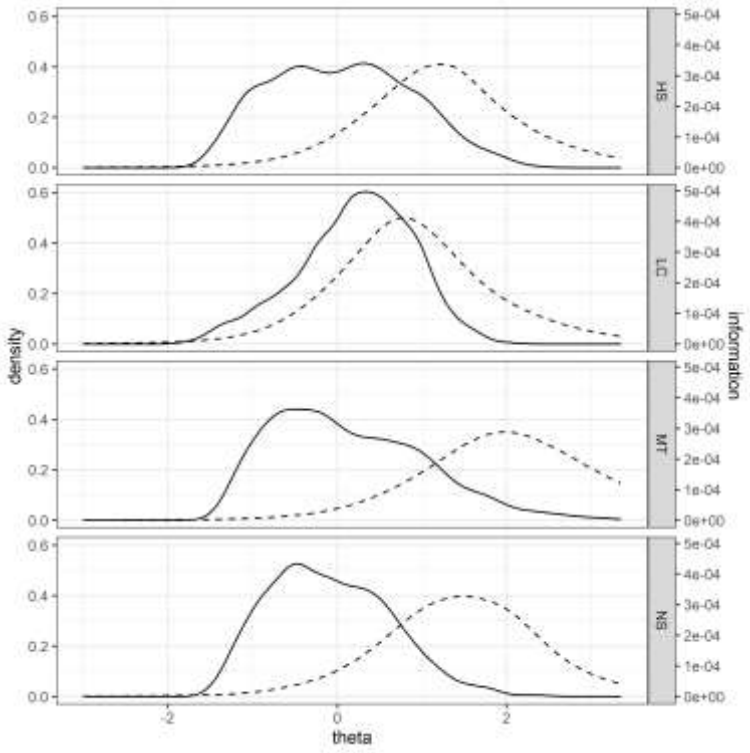
Table 4

Description of the four item banks

Area	n	mean (standard deviation)		
		a	b	c
HS	756	2,18 (0,95)	1,09 (0,80)	0,17 (0,07)
NS	749	2,34 (1,01)	1,39 (2,35)	0,17 (0,07)
LC	883	2,21 (0,9)	0,79 (0,82)	0,16 (0,07)
MT	792	2,09 (0,78)	1,97 (1,33)	0,16 (0,06)

Figure 3

Information curve of the item banks and density distribution of the thetas for each sample



Note. The solid line represents the density distribution of thetas for the sample, while the dashed line represents the item bank information.

The selection of the first item was random for conditions that used the random method or the PR method. For the MFI method, the initial theta was set to the mean of the scores from each area in the Enem 2020. For content balancing of the test, we used the modified constrained CAT method (MCCAT; Leung et al., 2000) as implemented in the simCAT package. This option restricts the test so that the same content is only administered after all other contents have been administered. We estimated theta using the EAP (Expected a Posteriori). The variable-length applications had a minimum of 15 items and a maximum of 60 items. The fixed-length applications had 20 (showed as a potential reduction by Jaloto & Primi, 2023b) and 45 (length of the traditional Enem) items. The simulation was conducted using the simCAT package (Jaloto & Primi, 2023a), which is inspired on the catR package (Magis & Raîche, 2012) but includes some differences, such as content balancing using MCCAT and stopping criterion based on error reduction.

CAT evaluation

The efficiency of variable-length CATs was evaluated in terms of the minimum, maximum, mean, and median values of application lengths. Lower values indicate greater reduction in test length provided by the CAT condition, thus making it more efficient. The precision of the CAT was evaluated based on the correlation, bias, and RMSE between the real theta and the simulated theta. Additionally, we assessed the standard error of measurement of the simulated theta. The real theta corresponded to the participant's official score in the transformed metric with a mean of 0 and standard deviation of 1, while the simulated theta corresponded to the score obtained in the simulation. In this study, we report the mean of these indicators across the 20 replications of the conditions.

Bias is a measure of the distance between the real theta and the simulated theta. It corresponds to the average difference between the real theta and the simulated theta, and is calculated as follows:

$$V = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_i\right)}{n} \tag{5}$$

where n is the total number of subjects, $\hat{\theta}_i$ is the estimated theta of subject i, and θ_i is the real theta of subject i. RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}} \tag{6}$$

The lower the RMSE, the more precise the CAT, as it indicates a smaller distance between the simulated theta and the real theta.

The security of the CAT was evaluated based on item exposure and overlap. Item exposure ranges from zero to one and is measured as the ratio between the number of times an item was administered and the number of subjects. An item with an exposure rate of zero was not administered. The higher the proportion of unadministered items, the more underutilized the item bank. An exposure rate greater than 0.30 indicates potential item overexposure. The higher the number of overexposed items, the less secure the item bank, as those items were

presented to a large proportion of participants. In addition to calculating item exposure rates, we established intervals of exposure rates and examined the number of items in each interval.

The overlap refers to the proportion of identical items that are presented to two randomly selected participants. The higher the overlap, the less secure the item bank is. It is calculated as follows (Chen et al., 2003)

$$\bar{T} = \frac{S^2 + \bar{r}}{\bar{r}} \quad (7)$$

where S^2 is the variance of the item exposure rates and $\bar{r}$ is the average of the item exposure rates.

In addition to reporting the overall average of the indicators in the 20 replications, we analyzed graphically the standard error and RMSE conditioned on theta for each area. For this graph, we also used the average of the 20 replications. Finally, we selected the condition with the best indicators of efficiency, precision, and security, and compared its precision with the precision of the linear test of Enem 2020.

Simulation of the linear test

With the response bank, we selected the responses to the items from the first administration of Enem 2020 and calculated the score of each participant in each test, using the same method as the theta estimation in CATs. Then, we compared the indicators of precision (standard error, bias, correlation, and RMSE) of the linear test with the indicators of the selected CAT from the previous step. Additionally, we compared the RMSE conditioned on theta for both tests.

Results

Eficiency of the CATs

Table 5 presents the efficiency indicators (average of minimum, maximum, mean, and median values of test length) for the eight variable-length CAT conditions in each area. Regarding the Standard error of 0.30 (SE30) stopping rule, in all four areas, the average test

length with the MFI method was smaller than with the progressive-restricted with acceleration parameters of 2 (PR2) and 3 (PR3) methods. In LC, the difference was less than two items, while in HS, NS, and MT, the difference exceeded five items, reaching up to 14 items. All three methods outperformed the random method, with differences of up to 30 items (NS, MIF). The average value of medians for the MFI method was 15 in HS, NS, and LC, indicating that in these areas, 15 items were sufficient to estimate the theta for 50% of the sample using this selection method. In MT, the median for this condition was 20. For the PR2 method, this value ranged from 15.0 (LC) to 40.1 (MT). For the PR3 method, the values ranged from 15.0 (LC) to 46.1 (MT). The median values for the random method were higher, reaching up to 60 in NS and MT. There were no major differences between the methods in terms of the average values of the minimum length (ranging from 15.0 to 15.4) and maximum length (60 in all conditions) of the tests in the four selection methods and areas.

The test lengths with the Standard error of 0.30 or Error reduction of 0.015 (ER015) stopping rule were smaller compared to SE30. The average values of the maximum test length did not exceed 23.5, and the average value of the median was 15 in all areas. This means that, on average, 50% of the sample responded to 15 items in all conditions with ER015. Additionally, the difference in average test length between the selection methods with ER015 was negligible and did not reach one item.

Table 5

Average of the minimum, maximum, mean, and median values of items administered in the simulations

Stopping rule	Selection method	HS				NS				LC				MT			
		Min	Max	NIA	MIA	Min	Max	NIA	MIA	Min	Max	NIA	MIA	Min	Max	NIA	MIA
SE30	RAN	15.0	60.0	45.1	56.6	15.0	60.0	50.8	60.0	15.0	60.0	33.5	27.2	15.4	60.0	54.6	60.0
	MFI	15.0	60.0	21.7	15.0	15.0	60.0	20.7	15.0	15.0	60.0	16.5	15.0	15.0	60.0	31.2	20.0
	PR2	15.0	60.0	26.9	19.0	15.0	60.0	29.6	24.0	15.0	60.0	19.6	15.0	15.0	60.0	40.5	40.1
	PR3	15.0	60.0	31.2	26.3	15.0	60.0	34.7	33.3	15.0	60.0	18.3	15.0	15.0	60.0	44.2	46.1
ER015	RAN	15.0	21.4	15.4	15.0	15.0	20.7	15.3	15.0	15.0	20.9	15.5	15.0	15.0	20.8	15.3	15.0

MFI	15.0	18.9	15.1	15.0	15.0	19.1	15.1	15.0	15.0	18.6	15.0	15.0	15.0	19.5	15.2	15.0
PR2	15.0	22.1	15.4	15.0	15.0	22.4	15.5	15.0	15.0	19.4	15.1	15.0	15.0	23.5	15.7	15.0
PR3	15.0	21.9	15.5	15.0	15.0	21.6	15.4	15.0	15.0	19.0	15.1	15.0	15.0	22.5	15.5	15.0

Note. HS = Human Sciences; NS = Natural Sciences; LC = Languages and Codes; MT = Mathematics; NIA = average number of items administered; Min = minimum; Max = maximum; MIA = median of items administered; RAN = random; MFI = Maximum Fisher Information; PR2 = progressive-restricted with $k = 2$; PR3 = progressive-restricted with $k = 3$; SE30 = standard error of 0.30; ER015 = standard error of 0.30 or error reduction of 0.015.

Precision of the CATs

Table 6 shows the precision indicators (mean of standard error, correlation, bias, and RMSE values) for the eight fixed-length CAT conditions in each area. In all areas, the MFI method showed the best values for these indicators, compared to the other selection methods within the same stopping rule. On the other hand, the random method showed the worst precision values. Despite the superiority of MFI, the difference between MFI and progressive-restricted (PR1 and PR2) methods was generally negligible. The highest mean standard error excluding the random method was 0.376, which corresponds to a reliability of 0.859. The lowest correlation among conditions excluding the random method was 0.926, the highest RMSE was 0.347, and the highest bias was 0.058. The average bias was positive in all conditions, indicating an overestimation of scores in the simulations. The largest differences in precision indicators were observed in MT and did not exceed 0.07 in all indicators. This indicates that fixed-length conditions with PR methods had satisfactory precision values.

Table 6

Average of standard error, correlation, bias, and root mean squared error of the replications of fixed-length CATs

Stopping rule	Selection method	HS				NS				LC				MT			
		SE	COR	Bias	RMSE	SE	COR	Bias	RMSE	SE	COR	Bias	RMSE	SE	COR	Bias	RMSE
FL45	RAN	0.370	0.918	0.053	0.340	0.423	0.870	0.066	0.372	0.271	0.929	0.045	0.268	0.510	0.880	0.078	0.439
	MFI	0.176	0.976	0.021	0.186	0.183	0.967	0.025	0.189	0.129	0.981	0.010	0.134	0.241	0.965	0.035	0.243
	PR1	0.200	0.969	0.024	0.208	0.224	0.953	0.031	0.227	0.189	0.960	0.021	0.197	0.288	0.954	0.042	0.281
	PR2	0.211	0.967	0.026	0.217	0.232	0.950	0.032	0.233	0.187	0.961	0.021	0.194	0.297	0.952	0.045	0.287

FL20	RAN	0.496	0.857	0.060	0.436	0.552	0.782	0.065	0.463	0.398	0.860	0.076	0.378	0.639	0.805	0.088	0.544
	MFI	0.240	0.960	0.029	0.239	0.242	0.948	0.032	0.238	0.185	0.964	0.019	0.187	0.311	0.947	0.046	0.299
	PR1	0.269	0.951	0.034	0.264	0.284	0.932	0.039	0.273	0.209	0.955	0.023	0.209	0.360	0.933	0.053	0.337
	PR2	0.285	0.947	0.035	0.275	0.300	0.926	0.042	0.285	0.218	0.952	0.024	0.217	0.376	0.928	0.058	0.347

Note. HS = Human Sciences; NS = Natural Sciences; LC = Languages and Codes; MT = Mathematics; SE = standard error; COR = correlation; RMSE = root mean square error; RAN = random; MFI = Maximum Fisher Information; PR1 = progressive-restricted with $k = 1$; PR2 = progressive-restricted with $k = 2$; FL20 = fixed-length (20 items); FL45 = fixed-length (45 items).

Table 7 shows the precision indicators for variable-length CATs. In these conditions, the MFI method also exhibited the best precision values, while the random method showed the worst. With the SE30 stopping criterion, the difference between MFI and progressive-restricted (PR2 and PR3) methods was negligible and did not exceed 0.06 for all precision indicators in all areas. With this stopping criterion, the highest average of standard error, excluding the random method, was 0.342, which corresponds to a reliability of 0.883. The lowest correlation among conditions excluding the random method was 0.921, the highest RMSE was 0.323, and the highest bias was 0.048. The mean bias was positive in all conditions.

Conditions with the ER015 stopping criterion had lower precision compared to SE30. Additionally, the differences between MFI and progressive-restricted (PR2 and PR3) methods were larger, reaching up to 0.29 in the mean standard error in MT. In HS, NS, and MT, PR3 behaved similarly to random. With the ER015 stopping criterion, the highest average of standard error, excluding the random method, was 0.628 (reliability of 0.606). The lowest correlation among conditions excluding the random method was 0.792, the highest RMSE was 0.541, and the highest bias was 0.088. The mean bias was positive in all conditions. In variable-length CATs, the precision of PR with the SE30 stopping criterion was satisfactory and similar to that observed in fixed-length CATs. However, with the ER015 stopping rule, the precision was unsatisfactory.

Table 7

Mean standard error, correlation, bias, and root mean square error of replications of variable-length CATs

Stopping rule	Selection method	HS				NS				LC				MT			
		SE	COR	Bias	RMSE	SE	COR	Bias	RMSE	SE	COR	Bias	RMSE	SE	COR	Bias	RMSE
SE30	RAN	0.375	0.918	0.048	0.338	0.404	0.881	0.054	0.360	0.313	0.911	0.034	0.293	0.481	0.894	0.076	0.418
	MFI	0.239	0.959	0.017	0.238	0.244	0.946	0.018	0.237	0.202	0.959	0.016	0.198	0.287	0.952	0.034	0.282
	PR2	0.274	0.947	0.019	0.270	0.286	0.925	0.015	0.278	0.240	0.943	0.020	0.233	0.333	0.940	0.043	0.317
	PR3	0.286	0.943	0.020	0.279	0.294	0.921	0.015	0.287	0.231	0.947	0.018	0.224	0.342	0.937	0.048	0.323
ER015	RAN	0.540	0.828	0.064	0.474	0.594	0.745	0.064	0.494	0.446	0.829	0.088	0.420	0.682	0.772	0.095	0.583
	MFI	0.268	0.952	0.032	0.261	0.264	0.940	0.034	0.257	0.211	0.955	0.022	0.209	0.334	0.941	0.048	0.316
	PR2	0.384	0.910	0.055	0.354	0.434	0.862	0.068	0.381	0.270	0.928	0.038	0.268	0.519	0.874	0.077	0.450
	PR3	0.469	0.868	0.061	0.420	0.538	0.792	0.072	0.453	0.254	0.936	0.031	0.251	0.628	0.808	0.088	0.541

Note. HS = Human Sciences; NS = Natural Sciences; LC = Languages and Codes; MT = Mathematics; SE = standard error; COR = correlation; RMSE = root mean square error; RAN = random; MFI = Maximum Fisher Information; PR2 = progressive-restricted with $k = 2$; PR3 = progressive-restricted with $k = 3$; SE30 = standard error of 0.30; ER015 = standard error of 0.30 or error reduction of 0.015.

RMSE and standard error conditioned on theta

Figure 4 shows the RMSE for each condition along the Enem scale. In fixed-length CATs, the use of exposure control had little impact on the RMSE across the entire scale. Additionally, the difference between PR1 and PR2 was practically nonexistent. Furthermore, MFI, PR1, and PR2 showed a significant gain in precision compared to the random method. The same pattern was observed in the variable-length CAT with the SE30 stopping rule: MFI, PR2, and PR3 showed a significant gain in precision, and exposure control did not significantly impact precision, with little difference between PR2 and PR3. However, in the CAT with the ER015 stopping rule, this pattern was only replicated in LC. In HS, NS, and MT, the gain in precision provided by MFI was not observed in PR2 and PR3. When exposure control was included, the RMSE in the lower regions of the scale (below zero) became similar to that of the random method. In the higher regions, although PR2 and PR3 improved precision, it remained well below the precision of MFI. Overall, Figure 4 shows that the precision with the use of PR as exposure control was adequate in fixed-length CATs and with the SE30 stopping rule.

Figure 4

RMSE conditioned on theta

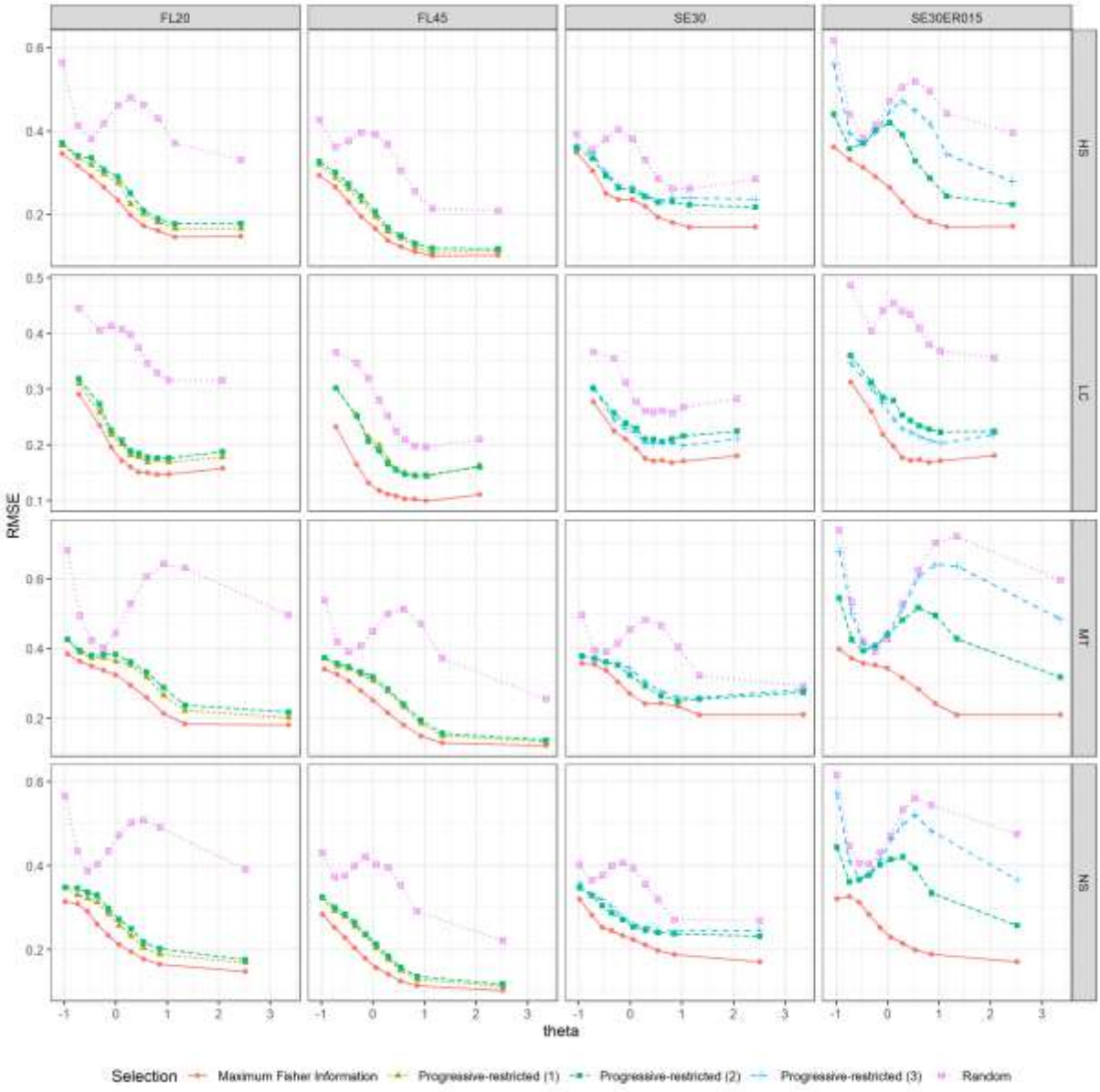
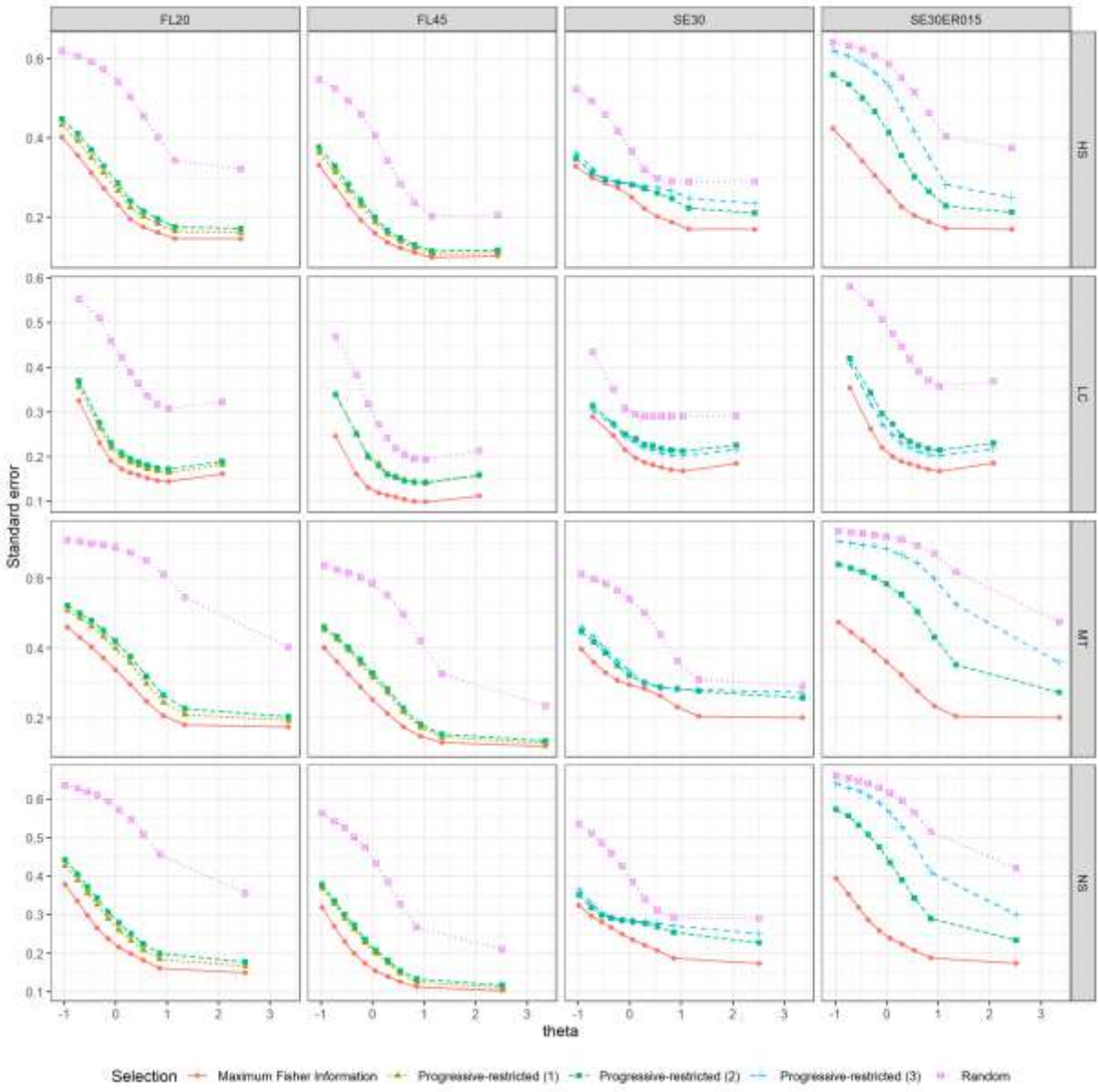


Figure 5 shows the standard error for each condition along the Enem scale. In fixed-length CATs, the general pattern of gain in standard error was similar to the pattern of gain in RMSE: MFI, PR1, and PR2 showed significant gains in precision, and exposure control had little impact on precision, with practically no difference between PR1 and PR2. In variable-length CATs, this pattern was also observed in LC. However, in HS, NS, and MT with the SE30 criterion, MFI, PR2, and PR3 showed significant gains in precision in the lower parts of the

scale with no major differences between these methods. Precision in the higher parts of the scale had a significant gain only with the MFI method, not with PR2 and PR3 methods. In these three areas, with the ER015 rule, precision with the MFI method showed significant gains across the entire scale. However, the gains with PR2 and PR3 methods were not significant in the lower parts of the scale, and precision remained unsatisfactory. Additionally, precision with PR3 was similar to precision with random selection. In the higher parts of the scale, PR2 showed significant gains, though not as high as MFI. The same pattern was observed with the PR3 method. Overall, Figure 5 shows that precision with the use of PR as exposure control was adequate in fixed-length CATs and with the SE30 stopping rule.

Figure 5

Standard error conditioned on theta



Security of CATs

The results regarding CAT security are not easily comparable across all conditions, unlike the results regarding precision. This is because larger tests naturally use more items in one administration, which increases the number of times an item is administered. Therefore, it is expected that the proportion of non-administered items will be lower. Additionally, an overlap rate of 0.30 in a 20-item test represents the administration of six common items between two subjects, whereas in a 45-item test, it represents 13.5 items. Therefore, we compare item exposures with caution across different CAT conditions.

Table 8 shows the security indicators of fixed-length CATs for each area (average of minimum and maximum exposure rates and overlap rate). In all areas, in conditions with random or PR2 method, all items were presented at least once, regardless of the test size. With the PR1 method, only in LC (FL20) there was a non-administered item. With the MFI method, there were non-administered items in all areas, with both test sizes.

As expected, the maximum exposure rates with the MFI method were 1.00. However, with exposure control, these rates were capped at a maximum of 0.30 (rounded). Some items exceeded this rate, but this is inherent to the exposure control method and occurs due to the way item filtering is performed at the beginning of the administration. With the restricted maximum information method, items that have more than 30% exposure are excluded and become available again after a few administrations, when their exposure drops below the threshold. The items that exceeded 30% exposure after the simulations are those that would not necessarily be available in the subsequent administrations, if there were any.

The item overlap rate decreased with the inclusion of exposure control. With the MFI method, it reached 0.432 (MT) with the FL45 criterion and 0.404 (MT) with the FL20 criterion. With PR2, the highest overlap rates were 0.151 (NS) with FL45 and 0.125 (MT) with FL20. The overlap rates for PR1 were slightly higher than those for PR2.

Table 8

Mean of the minimum and maximum exposure rates and item overlap rate in fixed-length CATs

Stopping rule	Selection method	HS			NS			LC			MT		
		r_{min}	r_{max}	O	r_{min}	r_{max}	O	r_{min}	r_{max}	O	r_{min}	r_{max}	O
FL45	RAN	0.037	0.092	0.061	0.038	0.089	0.062	0.014	0.098	0.056	0.035	0.098	0.059
	MFI	0.000	1.000	0.335	0.000	1.000	0.419	0.000	1.000	0.279	0.000	1.000	0.432
	PR1	0.004	0.300	0.174	0.005	0.300	0.182	0.002	0.300	0.155	0.005	0.300	0.181
	PR2	0.010	0.300	0.143	0.012	0.300	0.151	0.004	0.300	0.128	0.011	0.300	0.149
FL20	RAN	0.015	0.040	0.027	0.015	0.041	0.027	0.006	0.039	0.024	0.014	0.038	0.026
	MFI	0.000	1.000	0.323	0.000	1.000	0.383	0.000	1.000	0.264	0.000	1.000	0.404
	PR1	0.001	0.300	0.143	0.001	0.300	0.156	0.000	0.300	0.120	0.001	0.300	0.157
	PR2	0.002	0.300	0.110	0.003	0.300	0.122	0.001	0.299	0.091	0.003	0.300	0.125

Note. HS = Human Sciences; NS = Natural Sciences; LC = Languages and Codes; MT = Mathematics; r_{min} = minimum mean item exposure rate; r_{max} = maximum mean item exposure rate; O = mean overlap; RAN = random; MFI = Maximum Fisher Information; PR1 = progressive-restricted with $k = 1$; PR2 = progressive-restricted with $k = 2$; FL20 = fixed-length (20 items); FL45 = fixed-length (45 items).

Table 9 shows the security indicators of variable-length CATs for each area. In HS, NS, and MT, under conditions of random method or PR, all items were presented at least once, regardless of the stopping rule. In LC, there were unadministered items in both stopping rule even with exposure control. Similar to fixed-length CATs, MFI underutilized the item bank the most, as it had the highest number of items not presented. With the MFI method, there were unadministered items in all areas, using both stopping rule. As expected, the maximum item exposure rates with MFI method were 1.00. With exposure control, these rates were at most 0.30 (rounded). With the ER015 stopping rule, in HS, NS, and MT, the exposure control rates reached a maximum of 0.217 (HS).

The item overlap rate decreased with the inclusion of exposure control. With the MFI method, it reached 0.406 (MT) with both stopping rule. With PR2, the highest overlap rates were 0.167 (MT) with the SE30 criterion and 0.151 (LC) with ER015. The overlap rates of PR3 reached a maximum of 0.139 (MT) with the SE30 rule. With the ER015 stopping rule, the overlap values in HS, NS, and MT were similar to those of the random method.

Table 9

Mean of the minimum and maximum exposure rates and item overlap rate in variable-length CATs

Stopping rule	Selection method	HS			NS			LC			MT		
		r_{min}	r_{max}	O	r_{min}	r_{max}	O	r_{min}	r_{max}	O	r_{min}	r_{max}	O
SE30	RAN	0.036	0.096	0.061	0.040	0.105	0.070	0.009	0.075	0.042	0.041	0.131	0.072
	MFI	0.000	1.000	0.299	0.000	1.000	0.340	0.000	1.000	0.265	0.000	1.000	0.406
	PR1	0.005	0.300	0.119	0.007	0.300	0.127	0.000	0.300	0.148	0.007	0.300	0.167
	PR2	0.011	0.299	0.095	0.014	0.300	0.103	0.000	0.300	0.125	0.014	0.300	0.139
ER015	RAN	0.011	0.031	0.021	0.011	0.032	0.021	0.004	0.030	0.019	0.011	0.029	0.020
	MFI	0.000	1.000	0.335	0.000	1.000	0.389	0.000	1.000	0.284	0.000	1.000	0.406

PR1	0.004	0.217	0.047	0.006	0.177	0.042	0.000	0.300	0.151	0.006	0.127	0.033
PR2	0.008	0.109	0.025	0.010	0.073	0.023	0.000	0.300	0.125	0.009	0.045	0.021

Note. HS = Human Sciences; NS = Natural Sciences; LC = Languages and Codes; MT = Mathematics; r_{min} = minimum mean item exposure rate; r_{max} = maximum mean item exposure rate; O = mean overlap; RAN = random; MFI = Maximum Fisher Information; PR2 = progressive-restricted with $k = 2$; PR3 = progressive-restricted with $k = 3$; SE30 = standard error of 0.30; ER015 = standard error of 0.30 or error reduction of 0.015.

Table 10 shows the mean percentage of items per exposure rate interval in fixed-length applications. To obtain this value, we calculated the average exposure rates for each item in the 20 replications and verified the number of items in each established interval. For both test lengths, the MFI method had the highest proportions of items not administered (reaching up to 82.6% in MT with FL20) and items overexposed (reaching up to 7.5% in HS and NS with FL45). As expected, the inclusion of exposure control reduced the proportion of items not presented and overexposed. The average of items not administered with PR1 and PR2 methods in all areas and with both test lengths was zero. A few items had exposure rates higher than 0.30, but this is inherent to the exposure control method, as explained earlier. Regarding to the acceleration parameter, the higher its value, the lower the proportion of items with higher exposure rates for both test lengths.

Table 10

Overall average percentages of items for each exposure rate interval in fixed-length CATs

		HS				NS				LC				MT			
Stopping rule	Exposure	RAN	MFI	PR1	PR2	RAN	MFI	PR1	PR2	RAN	MFI	PR1	PR2	RAN	MFI	PR1	PR2
FL20	0	0.0	79.8	0.0	0.0	0.0	80.2	0.0	0.0	0.0	74.5	0.0	0.0	0.0	82.6	0.0	0.0
	(0;0.02]	0.0	6.7	78.7	76.1	0.0	7.7	80.6	78.8	24.3	9.7	80.1	77.3	0.0	5.8	81.6	80.4
	(0.02;0.05]	100.0	2.4	9.3	12.8	100.0	2.7	8.0	11.3	75.7	5.0	8.4	11.9	100.0	2.7	7.4	9.2
	(0.05;0.1]	0.0	2.1	4.4	5.0	0.0	1.6	3.9	3.3	0.0	2.0	4.5	5.2	0.0	2.0	3.5	4.4
	(0.1;0.15]	0.0	2.4	2.5	2.1	0.0	1.6	1.9	2.0	0.0	3.5	3.5	2.7	0.0	1.5	2.1	1.5
	(0.15;0.2]	0.0	1.3	1.7	1.7	0.0	0.8	1.3	1.5	0.0	1.2	1.4	1.6	0.0	1.3	1.5	1.4
	(0.2;0.25]	0.0	1.3	1.5	0.9	0.0	0.9	1.6	1.3	0.0	1.2	0.9	1.0	0.0	0.6	0.8	1.3
	(0.25;0.3]	0.0	1.5	1.6	1.3	0.0	1.2	2.7	1.7	0.0	1.0	1.2	0.2	0.0	0.3	2.5	1.6
	(0.3;0.4]	0.0	1.2	0.4	0.0	0.0	1.1	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.9	0.5	0.1
	(0.4;1]	0.0	1.3	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.8	0.0	0.0	0.0	2.4	0.0	0.0
		HS				NS				LC				MT			
	Exposure	RAN	MFI	PR1	PR2	RAN	MFI	PR1	PR2	RAN	MFI	PR1	PR2	RAN	MFI	PR1	PR2

FL45	0	0.0	59.1	0.0	0.0	0.0	60.2	0.0	0.0	0.0	57.1	0.0	0.0	0.0	62.4	0.0	0.0
	(0;0.02]	0.0	12.8	48.5	21.3	0.0	15.8	53.4	11.9	2.3	13.6	52.9	35.9	0.0	13.6	55.2	18.6
	(0.02;0.05]	11.6	4.9	22.8	50.4	13.2	4.4	20.4	62.1	37.9	5.2	20.2	38.7	25.8	5.1	19.8	57.8
	(0.05;0.1]	88.4	4.6	11.0	13.0	86.8	3.2	7.9	11.5	59.8	6.7	11.9	13.3	74.2	5.2	8.8	10.4
	(0.1;0.15]	0.0	3.4	4.1	4.9	0.0	2.5	4.8	3.7	0.0	4.1	4.6	3.9	0.0	1.8	3.3	3.2
	(0.15;0.2]	0.0	3.2	3.6	2.8	0.0	2.9	2.7	2.8	0.0	2.8	3.3	2.9	0.0	1.6	2.3	2.1
	(0.2;0.25]	0.0	2.0	2.6	2.1	0.0	1.9	2.4	1.2	0.0	3.7	2.2	1.9	0.0	1.8	2.3	1.5
	(0.25;0.3]	0.0	2.4	5.8	4.4	0.0	1.6	8.4	6.8	0.0	1.9	4.6	3.2	0.0	1.3	5.7	4.4
	(0.3;0.4]	0.0	3.4	1.6	1.2	0.0	1.9	0.0	0.0	0.0	2.6	0.3	0.2	0.0	2.0	2.7	2.0
	(0.4;1]	0.0	4.1	0.0	0.0	0.0	5.6	0.0	0.0	0.0	2.3	0.0	0.0	0.0	5.3	0.0	0.0

Note. HS = Human Sciences; NS = Natural Sciences; LC = Languages and Codes; MT = Mathematics; RAN = random; MFI = Maximum Fisher Information; PR1 = progressive-restricted with $k = 1$; PR2 = progressive-restricted with $k = 2$; FL20 = fixed-length (20 items); FL45 = fixed-length (45 items).

Table 11 shows the overall average percentages of items for each exposure rate interval in variable-length CATs. For both stopping rules, the MFI method had higher proportions of items not administered (reaching up to 87.4% in MT with ER015) and items overexposed (up to 5.2% in MT with SE30). Similar to fixed-length CATs, the inclusion of exposure control reduced the proportion of items not administered and overexposed. The average percentage of items not administered in all four areas with PR2 or PR3 methods for both stopping rules was zero. Additionally, few items had exposure rates higher than 0.30 with PR2 or PR3, as explained previously. Regarding the acceleration parameter, higher values led to lower proportions of items with higher exposure rates for both stopping rules.

Table 11

Overall average percentages of items for each exposure rate interval in variable-length CATs

Stopping rule	Exposure	HS				NS				LC				MT			
		RAN	MFI	PR2	PR3	RAN	MFI	PR2	PR3	RAN	MFI	PR2	PR3	RAN	MFI	PR2	PR3
SE30	0	0.0	75.1	0.0	0.0	0.0	76.4	0.0	0.0	0.0	71.2	0.0	0.0	0.0	78.5	0.0	0.0
	(0;0.02]	0.0	8.2	65.7	18.8	0.0	9.1	64.5	0.8	8.3	15.2	80.1	79.8	0.0	6.7	52.4	1.9
	(0.02;0.05]	16.0	3.3	18.3	64.7	0.3	2.0	20.2	81.8	80.2	4.8	8.2	9.2	0.4	2.4	28.2	78.7
	(0.05;0.1]	84.0	3.7	7.1	8.6	99.7	3.7	5.2	8.1	11.6	3.1	5.2	5.4	96.8	1.5	5.9	6.8
	(0.1;0.15]	0.0	2.5	2.6	2.9	0.0	2.4	2.7	3.3	0.0	1.9	2.0	2.4	2.8	1.1	2.7	3.3
	(0.15;0.2]	0.0	1.9	2.5	2.2	0.0	1.9	3.2	2.4	0.0	0.8	1.5	1.0	0.0	0.9	2.4	2.0
	(0.2;0.25]	0.0	1.9	1.7	1.5	0.0	1.1	2.0	1.5	0.0	0.9	1.0	1.0	0.0	1.6	1.4	1.9
	(0.25;0.3]	0.0	0.8	2.0	1.3	0.0	0.8	2.3	2.0	0.0	0.8	1.9	1.1	0.0	2.0	5.7	4.5
	(0.3;0.4]	0.0	1.2	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.6	0.1	0.0	0.0	1.8	1.4	0.9

	(0.4;1]	0.0	1.5	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.8	0.0	0.0	0.0	3.4	0.0	0.0
				HS				NS				LC				MT	
	Exposure	RAN	MFI	PR2	PR3	RAN	MFI	PR2	PR3	RAN	MFI	PR2	PR3	RAN	MFI	PR2	PR3
	0	0.0	85.3	0.0	0.0	0.0	85.0	0.0	0.0	0.0	81.4	0.0	0.0	0.0	87.4	0.0	0.0
	(0;0.02]	39.2	4.0	74.3	70.0	41.8	6.0	76.0	68.5	67.0	7.8	86.3	85.1	69.4	4.0	76.0	70.6
	(0.02;0.05]	60.8	2.1	18.9	27.6	58.2	1.6	18.0	30.3	33.0	2.4	4.8	6.1	30.6	2.0	17.6	29.4
ER015	(0.05;0.1]	0.0	2.1	5.0	2.1	0.0	1.2	3.9	1.2	0.0	2.6	3.6	4.5	0.0	1.0	5.6	0.0
	(0.1;0.15]	0.0	1.5	1.1	0.3	0.0	1.7	1.6	0.0	0.0	2.0	1.8	1.2	0.0	1.3	0.9	0.0
	(0.15;0.2]	0.0	1.2	0.5	0.0	0.0	0.5	0.5	0.0	0.0	0.7	0.8	1.1	0.0	1.1	0.0	0.0
	(0.2;0.25]	0.0	0.9	0.1	0.0	0.0	0.7	0.0	0.0	0.0	0.9	1.2	1.0	0.0	0.3	0.0	0.0
	(0.25;0.3]	0.0	1.1	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.8	1.5	0.9	0.0	0.5	0.0	0.0
	(0.3;0.4]	0.0	0.7	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.6	0.0	0.0	0.0	1.0	0.0	0.0
	(0.4;1]	0.0	1.2	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.8	0.0	0.0	0.0	1.4	0.0	0.0

Note. HS = Human Sciences; NS = Natural Sciences; LC = Languages and Codes; MT = Mathematics; RAN = random; MFI = Maximum Fisher Information; PR2 = progressive-restricted with $k = 2$; PR3 = progressive-restricted with $k = 3$; SE30 = standard error of 0.30; ER015 = standard error of 0.30 or error reduction of 0.015.

In general, under the same stopping rule conditions, the random method provided the safest test administrations, while the MFI method was the least secure. The inclusion of the PR exposure control method significantly increased test security, and higher acceleration parameter values resulted in greater test security.

Comparison between CAT and linear

The most satisfactory combination of CAT was the fixed length of 20 items with the PR2 method (PR2FL20). In terms of efficiency, the average length with ER015 was less than 16 in all areas. Additionally, some variable-length CAT conditions had a maximum length of less than 20 items. However, this decrease in test length was in no more than five items, and in terms of precision, FL20 CATs outperformed variable-length CATs, especially those with ER015. Therefore, the slightly higher efficiency of the ER015 rule does not justify the loss in precision compared to FL20. In parallel, the use of exposure control significantly increased the test's security without compromising application precision. Therefore, the condition adopted for comparison with the linear application of Enem was PR2FL20.

Table 12 presents the precision indicators of the linear test simulation. Additionally, this table reiterates the values of CATs with the PR2FL20 condition for ease of comparison. The

precision obtained in the linear test was satisfactory, as the highest average standard error was 0.495 (MT), which corresponds to a reliability of 0.755, the lowest average correlation was 0.860 (NS), and the highest value of RMSE was 0.433 (MT). The highest absolute bias was 0.078 (NS). However, in all areas, the precision indicators were better in the CAT with PR2FL20.

Table 12

Mean of standard error, correlation, bias, and root mean square error of the replications of simulations with linear application and PR2FL20 condition of the CAT.

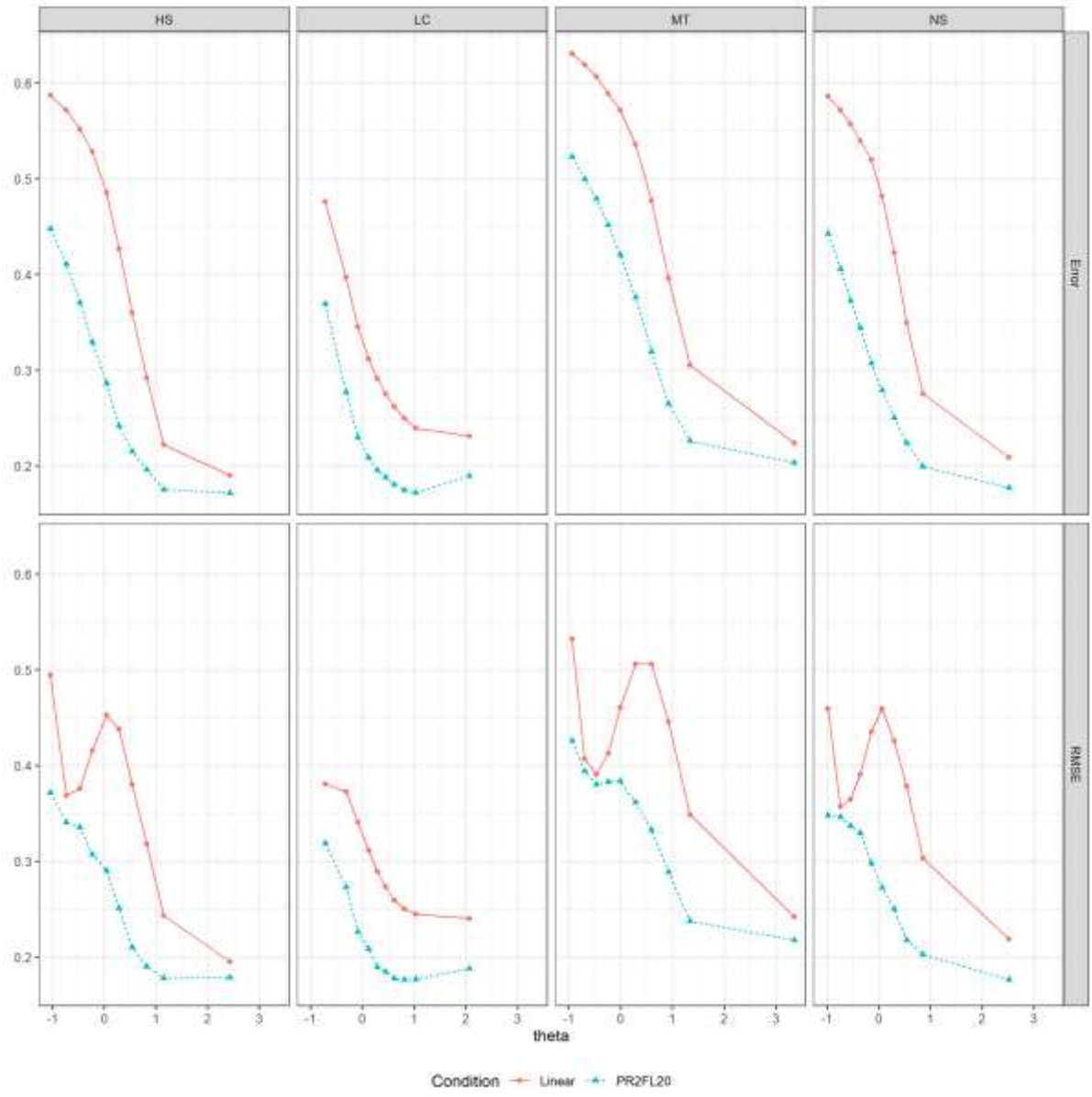
	Linear				PR2FL20			
	Standard Error	Correlation	Bias	RMSE	Standard Error	Correlation	Bias	RMSE
HS	0,422	0,895	-0,061	0,379	0,285	0,947	0,035	0,275
NS	0,451	0,860	-0,078	0,386	0,300	0,926	0,042	0,285
LC	0,308	0,908	-0,052	0,301	0,218	0,952	0,024	0,217
MT	0,495	0,885	-0,077	0,433	0,376	0,928	0,058	0,347

Note. PR2FL20 = progressive-restricted with $k = 2$ and fixed-length (20 items); RMSE = root mean square error.

Figure 6 shows the RMSE and standard error of the linear application along the Enem scale, as well as the CAT with PR2FL20. Throughout the entire scale, the CAT shows a significant improvement in RMSE compared to the linear application. The same trend is observed for the standard error.

Figure 6

Standard error and RMSE conditioned on theta for linear test and CAT with PR2FL20



Note. PR2FL20 = progressive-restricted with $k = 2$ and fixed-length (20 items); RMSE = root mean square error.

Discussion

The objective of this study was to evaluate the exposure control PR with different acceleration parameters in fixed and variable length CATs. We simulated the administration of Enem (a high-stakes educational test administered on paper) in CAT format. The results partially support hypothesis H1 (the efficiency and precision of PR method will be similar to MFI). The efficiency of PR method was lower than MFI in variable-length CATs, as the number of items administered with PR method was higher. On the other hand, precision with PR was slightly lower than with MFI in fixed-length CATs and CAT with SE30 stopping rule. With

ER015 stopping rule, precision had a relevant negative impact. Hypothesis H2 (the security of PR will be higher than MFI) was corroborated. Hypothesis H3 (the efficiency and precision of PR method will be similar with all acceleration parameters) was partially supported. In fixed-length CATs and CAT with SE30 stopping rule, changes in acceleration parameter did not have a relevant impact on precision. In the latter stopping rule, efficiency reduced with increased acceleration parameter. With ER015 stopping rule, efficiency was not strongly affected, but precision had a negative impact. Hypotheses H4 (higher acceleration parameter leads to higher security) and H5 (precision of CAT with PR will be higher than linear test) were corroborated. Specifically, regarding H5, we compared the precision of linear test with that of a fixed length CAT with 20 items and PR method with an acceleration parameter of 2.

Our results regarding the efficiency of CAT with SE30 stopping rule contrast with previous findings. In this study, the average number of items administered increased considerably in conditions with PR, with differences of up to 14 items compared to MFI. In Leroux et al.'s (2013) study, the difference reached a maximum of seven items. In Leroux and Dodd's (2016) study, the difference did not exceed three items, and in Leroux et al.'s (2019) study, the difference in the average number of items administered did not even reach one item. The fact that the difference in test length is larger in this study may be due to the characteristics of the item bank and the sample. Leroux and Dodd (2016) and Leroux et al.'s (2019) studies used polytomous items, which may widen the ability range where the test provides satisfactory total information, compared to a test with dichotomous items. As a result, the random administration of items may have a smaller impact, as the selected item is likely to contribute significantly to the reduction of measurement error. This hypothesis is reinforced by the fact that Leroux et al. (2013), who used dichotomous items, found a larger average difference than the other studies. Another possible cause of the difference between our findings and previous ones is the use of the acceleration parameter. In this study, the higher the acceleration parameter,

the larger the average test length. We did not use PR1 in the variable-length CATs, but in conditions with PR2, the maximum difference from MFI was nine items, which is closer to Leroux et al.'s (2013) findings. It should be noted that caution is needed when comparing findings, as the formula used by these previous studies for item selection in PR (McClarty et al., 2006) does not include an acceleration parameter and would approximate Equation 3 with an acceleration parameter of 1.

We found a negligible reduction in CAT precision when including PR exposure control in fixed-length CATs and with a stopping rule based on standard error, corroborating previous findings (Barrada et al., 2008; Leroux & Dodd, 2016; Leroux et al., 2013, 2019). However, the use of PR method with the stopping rule of error reduction had a relevant impact on measurement error. This result contrasts with Leroux et al.'s (2019) findings, who used the predicted standard error reduction (PSER) stopping rule. The difference adopted as a criterion in the cited study was 0.020, which is less stringent than the one used in this study. However, in PSER, even if the standard error reaches the desired value, the application continues until the error reduction reaches a determined value. In contrast, in the conditions of this study with the error reduction criterion, the application would stop if the standard error reached 0.30. Therefore, a possible reason for the differences in results between the studies is the reduction of the mean standard error caused by this continuous application. Hence, we recommend studies that compare the accuracy of cases where the application is terminated with a standard error greater than the desired value, as well as the accuracy of cases with a standard error lower than the desired value.

Another possible reason for the decrease in accuracy when combining PR with the error reduction rule is early termination of the application. The weight of the random component of the PR method decreases more slowly with higher acceleration parameter values. Given the random selection of items, two consecutive items with low informativeness for the provisional

theta may be administered, even after reaching the minimum test length (in the case of this study, 15 items). If this occurs, there may be a low reduction in error. If this reduction is lower than the stopping criterion (in our case, 0.015), the application terminates prematurely. In the case of PSER, the algorithm checks the potential reduction in measurement error for each available item in the item bank and subsequently selects the item to be administered. This difference in the algorithm may mitigate the effect of PR randomness. Therefore, we recommend studies that investigate the impact of PR randomness on the early termination of CATs with the error reduction stopping rule. Additionally, we suggest that studies explore minimum test lengths or algorithms that prevent premature termination of the application.

Our findings on test security corroborate previous research. We observed a reduction in item overlap rate when using the PR method, and this rate increased as the acceleration parameter also increased. These results were also found by Barrada et al. (2008) in fixed-length CATs. We did not find studies that evaluated the effect of the acceleration parameter on test security in variable-length CATs, so we recommend replicating the simulation conditions used in this study with other item banks and subjects to further investigate this aspect.

The differences in efficiency, precision, and security indicators across the four tests reinforce that their values are impacted by the distribution of items and thetas along the scale. Overall, the indicators were better for the LC test, whose test information curve most closely matched the density curve of the sample theta. This finding is in line with Lee and Dodd's (2012) study, which simulated CATs with item banks of different distributions (one with a peak at the easy end of the scale, one at the medium range, and one at the difficult end) and two groups of subjects (one with a mean of zero and the other with a mean of 0.74). The best precision and security results were obtained with the medium-difficulty item bank, regardless of the subject group. The worst results were obtained with the easy item bank and the group of

subjects with higher mean. This underscores the importance of investigating the effect of item bank distribution beyond the CAT algorithm.

The reduction of the length of the Enem test in this study to 20 items was greater than the reduction to 33 items proposed by Spenassato et al. (2016) for the MT test from Enem 2012. The increase in efficiency in this study was expected, as the MT item bank consisted of 792 items, whereas the cited study used 45 items. The authors found an average standard error of 0.351, while in this study, this measure was 0.376. This difference may be due to the use of exposure control, as the average standard error in MT in the condition with MFI in this study was 0.311. This means that the significantly larger size of our item bank likely compensated for the impact of reducing the length of the test and the exposure control in CAT Enem, as the difference in error compared to the previous study was small. Although our RMSE and correlation values (0.347 and 0.928) were less satisfactory than those of the previous study (0.088 and 0.998), this can be explained by the fact that the authors considered the theta estimated with the 45 items in their study as the true theta. Since 33 items were administered, it makes sense that the estimated theta in the CAT would be very close to the true theta. In this study, instead of calculating the true theta from the responses, we simulated the responses from the considered true theta.

The reduction observed in this study was similar to that observed by Jatobá et al. (2020), who also used the 45 items from the MT test of Enem 2012. In that study, the authors used a customized item selection method, which allowed for a reduction to 21 items, while the application with MFI reduced to 35 items. In this study, we achieved a length of 20 items using a universal item selection method, so future studies should investigate the potential of other selection methods to further increase the efficiency of CAT Enem.

Our reduction was also greater than the one observed by Tabak et al. (2023), who implemented a 25-tem multistage adaptive test with the 45 MT items from Enem 2019.

However, their indicators of precision were relative better than ours (correlation of 0.967 and RMSE of 0.12), which show the potentiality of applying multistage testing in Enem. Therefore, we encourage studies that implements this selection method with large item banks and exposure control.

Our reduction also exceeded that of Kalender and Berberoglu (2017). The fixed-length CAT of 25 items in their study had average standard errors ranging from 0.25 to 0.32. In this work, the 20-item CATs with MFI had an average standard error of up to 0.311. Considering that the authors' item bank had 45 items, this improvement was expected. In the CAT that we selected as the most satisfactory, which controlled item exposure, the average standard error ranged from 0.218 to 0.376. This represents a negligible loss in precision, considering the gain in test security for a high-stakes examination.

This study has the limitation of using only simulated responses. Therefore, the responses were not subject to possible external influences (such as fatigue, motivation, and prior knowledge of the participant about an item). We recommend further studies that compare the precision and efficiency of CATs with linear tests in high-stakes contexts. We highlight the possibility of developing an efficient CAT for the Enem that can improve its precision and security. Considering the four tests, the CAT selected in this study totaled 80 items, which is a smaller quantity than the total number of items in a single day of exam administration. We hope to contribute to optimizing high-stakes educational tests, particularly the Enem, in a way that makes the measurement process fairer by eliminating undesirable interference.

Acknowledgments

We thank Inep researcher Giordano Sereno for clarifying doubts about the microdata.

Competing interest

Alexandre Jaloto is researcher at the National Institute for Educational Studies and Researches Anísio Teixeira (Inep). The opinions expressed are those of the authors and do not represent the views of the institute or the Brazilian Ministry of Education.

Disclosure statement

The opinions expressed in this publication are exclusively and full responsibility of the authors, not necessarily expressing the point of view of Inep or the Brazilian Ministry of Education.

Authors' contributions

Alexandre Jaloto designed the study, conducted the analyses, and drafted the manuscript. Ricardo Primi contributed to the interpretation of the results and reviewed the manuscript. All authors reviewed and approved the final manuscript.

References

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology, 61*(2), 493–513. 10.1348/000711007X230937

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, (49), 61–80.

Chen, S.-Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*(2), 129–145.

Jaloto, A., & Primi, R. (2023a). *SimCAT. Computerized adaptive testing simulations*. https://github.com/alexandrejaloto/simCAT

Jaloto, A., & Primi, R. (2023b). Next-generation Enem assessment with fewer items and high reliability using CAT. *SciELO Preprints*. https://doi.org/10.1590/SciELOPreprints.5339

Jatobá, V. M. G., Farias, J. S., Freire, V., Ruela, A. S., & Delgado, K. V. (2020). ALICAT: A customized approach to item selection process in computerized adaptive testing. *Journal of the Brazilian Computer Society, 26*(1), 4. 10.1186/s13173-020-00098-z

Kalender, I., & Berberoglu, G. (2017). Can computerized adaptive testing work in students' admission to higher education programs in Turkey?. *Educational Sciences: Theory & Practice, 17*(2), 573–596. http://doi.org/10.12738/estp.2017.2.0280

Kallen, M. A., Cook, K. F., Amtmann, D., Knowlton, E., & Gershon, R. C. (2018). Grooming a CAT: Customizing CAT administration rules to increase response efficiency in

specific research and clinical settings. *Quality of Life Research*, *27*(9), 2403–2413. 10.1007/s11136-018-1870-z

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement*, *72*(1), 159–175. 10.1177/0013164411411296

Leroux, A. J., & Dodd, B. G. (2016). A comparison of exposure control procedures in CATs using the GPC model. *The Journal of Experimental Education*, *84*(4), 666–685. 10.1080/00220973.2015.1099511

Leroux, A. J., Lopez, M., Hembry, I., & Dodd, B. G. (2013). A comparison of exposure control procedures in CATs using the 3PL model. *Educational and Psychological Measurement*, *73*(5), 857–874. 10.1177/0013164413486802

Leroux, A. J., Waid-Ebbs, J. K., Wen, P.-S., Helmer, D. A., Graham, D. P., O'Connor, M. K., & Ray, K. (2019). An investigation of exposure control methods with variable-length CAT using the partial credit model. *Applied Psychological Measurement*, *43*(8), 624–638. 10.1177/0146621618824856

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2000). Content balancing in stratified computerized adaptive testing designs. *Annual Meeting of the American Educational Research Association*, New Orleans. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1058.3442&rep=rep1&type=pdf

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: recent updates of the package catR. *Journal of Statistical Software*, *76*(Code Snippet 1). 10.18637/jss.v076.c01

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8). 10.18637/jss.v048.i08

McClarty, K. L., Sperling, R. A., & Dodd, B. G. (2006). A variant of the progressive-restricted item exposure control procedure in computerized adaptive testing systems based on the 3PL and partial credit models. *Annual Meeting of the American Educational Research Association*, San Francisco

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the Word Part Levels Test. *Language Testing*, *36*(1), 101–123. 10.1177/0265532217725776

Morris, S. B., Bass, M., Howard, E., & Neapolitan, R. E. (2020). Stopping rules for computer adaptive testing when item banks have nonuniform information. *International Journal of Testing*, *20*(2), 146–168. 10.1080/15305058.2019.1635604

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(4), 311–327.

Spenassato, D., Trierweiller, A. C., Andrade, D. F. de, & Bornia, A. C. (2016). Testes Adaptativos Computadorizados Aplicados em Avaliações Educacionais. *Revista Brasileira de Informática na Educação*, *24*(02), 1. http://doi.org/10.5753/rbie.2016.24.02.1

Sulak, S., & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *10*(3), 315–326. 10.21031/epod.530528

Tabak, G. C., Piton-Gonçalves, J., Ricarte, T. A. M., & Curi, M. (2023). Teste Adaptativo Multiestágio para o ENEM. *Revista Brasileira de Informática na Educação*, *31*, 60–86. 10.5753/rbie.2023.2529

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

This preprint was submitted under the following conditions:

- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.