Publication status: Preprint has been submitted for publication in journal

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: 2023-12-11 (version 3)
(YYYY-MM-DD)

Version justification: We have reduced the number of pages to comply with the journal's guidelines.

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

Alexandre Jaloto[a]* and R. Primi[b]

[a]*Inep, Brasília, Brazil;* [b]*Department of Psychology, Universidade São Francisco, Campinas, Brazil*

alexandre.jaloto@inep.gov.br

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

Subject classification codes: include these here if the journal requires them

Introduction

Computerized Adaptive Testing (CAT) can increase test efficiency by reducing its length (Bulut & Kan, 2012; Mizumoto et al., 2019; Spenassato et al., 2016). However, in high-stakes tests, security (in terms of item exposure) must also be considered, alongside efficiency. Through a random component, the progressive-restricted (PR) method manages item exposure without compromising efficiency or 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 used for higher education admission.

CAT optimizes test administration because items are administered to the participant based on their response to the previous item (Weiss & Kingsbury, 1984). Maximum Fisher Information (MFI) method, selecting items with the highest information for the provisional theta (latent variable) enhances efficiency (in terms of test length) without compromising 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.

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. 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., 2013, 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:

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

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 in fixed-length CAT 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.

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

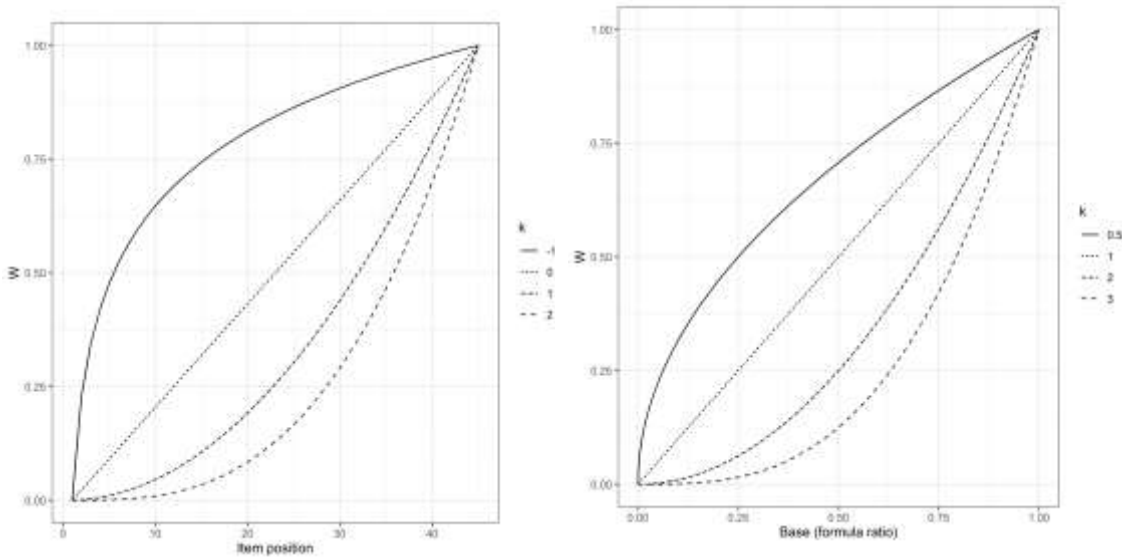
$$W = \begin{cases} 0, & \text{if } t = 0 \\ max\left[\frac{I(\hat{\theta}_t)}{I_{stop}}, \frac{q}{M-1}\right]^k, & \text{if } t \neq 0 \end{cases} \quad (3)$$

where I_{stop} is the information needed to reach the stopping value of the standard error and M is the maximum test length. Figure 1 illustrates four situations of k values in a 45-item CAT and in a variable-length CAT. For the latter, the x-axis corresponds to the formula ratio, which is the base in Equation 3, i.e., $max\left[\frac{I(\hat{\theta}_t)}{I_{stop}}, \frac{q}{M-1}\right]$.

From Figure 1, 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.

Figure 1. Variation of the weight in the progressive-restricted method with different acceleration parameters, in fixed-length and variable-length CATs



A) fixed-length CAT

B) variable-length CAT (x-axis corresponds to the base in Equation 3)

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 Brazilian Institute for

Educational Studies and Researches Anísio Teixeira (Inep), is composed of an essay and four tests with 45 multiple-choice items each, namely: Human Sciences (HS), Natural Sciences (NS), Languages and Codes (LC), and Mathematics (MT). For this paper, we used the MT test.

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 a robust item bank (792 items) previously administered in a real-world setting.

Methods

Study design

For each type of CAT (fixed-length and variable- length), we manipulated 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.

We had 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). The error reduction stopping rule performed well in previous simulations (e.g., Kallen et al., 2018; Morris et al., 2020). As we replicated the response bank 20 times, this study had a total of 320 simulations. Table 1 shows the conditions of the simulations. 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			
Fixed-length CATs				
	Random	MFI	PR with AP 1	PR with AP 2
Fixed length of 45 items (FL45)	RANFL45	MFIFL45	PR1FL45	PR2FL45
Fixed length of 20 items (FL20)	RANFL20	MFIFL20	PR1FL20	PR2FL20
Variable-length CATs				
	Random	MFI	PR with AP 2	PR with AP 3
Standard error of 0.30 (SE30)	RANSE30	MFISE30	PR2SE30	PR3SE30
Standard error of 0.30 or Error reduction of 0.015 (ER015)	RANER015	MFIER015	PR2ER015	PR3ER015

Note. MFI = Maximum Fisher information PR = progressive-restricted; AP = acceleration parameter of progressive method.

Response bank

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. By adopting this procedure, we were able to generalize our results to the population of the 2020 edition of the Enem. Additionally, we supplemented the sample size until it was three times larger than the number of items in the item bank. This allows other researchers to compare their results related to item exposure with ours more robustly, as one may keep the ratio between the size of the item bank and the size of the participants constant. The descriptive statistics of the participants in the Enem 2020 and the sample are presented in Table 2.

Table 2. Descriptive statistics of the participants in the Enem 2020 and the simulation samples

	N	Mean (standard deviation)	Range
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

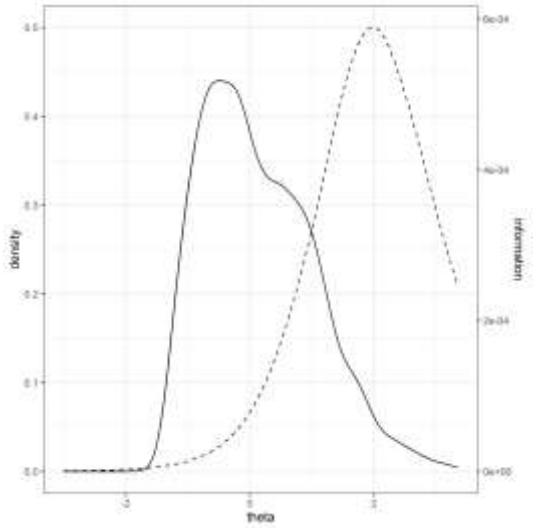
Item responses were generated through Monte Carlo simulations, creating a response matrix where rows represented subjects and columns represented the 792 items. This approach simulated responses as if each subject had answered all items, ensuring robust replication. We replicated the simulation 20 times.

CAT specifications

The item bank was composed of items administered in the Enem from 2009 to 2020, obtained from the microdata in November 2022. 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). Item discrimination had a mean of 2.09 (SD = 0.78), item difficulty had a mean of 1.97 (SD = 1.33) and item pseudoguess had a mean of 0.16 (SD = 0.06). Figure 2 shows the test information curve and the density distribution of the thetas for the sample.

Figure 2. 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.

For the MFI method, the initial theta was set to the mean of the scores in the Enem 2020. For content balancing of the test, we used the modified constrained CAT method (MCCAT; Leung et al., 2000). 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

Efficiency of variable-length CATs was assessed based on minimum, maximum, mean, and median application lengths, with lower values indicating greater test length reduction, enhancing efficiency. Precision evaluation involved correlation, bias, and RMSE between real and simulated thetas, with real theta representing participants' official scores. Standard error of measurement for simulated theta was also considered. Mean values of these indicators across 20 replications were reported.

Bias, measuring the difference between real and simulated thetas, was calculated as

$$V = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_i\right)}{n} \tag{4}$$

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. RMSE was calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}} \tag{5}$$

The lower the RMSE, the more precise the CAT.

Security evaluation considered item exposure and overlap. Item exposure, ranging from zero to one, represented the ratio of item administrations to subjects. An exposure rate over 0.30 indicated potential item overexposure. Overexposed items, administered to a large proportion of participants, reduced item bank security. In addition to calculating item exposure rates, we established intervals of exposure rates and examined the number of items in each interval.

Overlapping items presented to two randomly chosen participants were assessed using (Chen et al., 2003)

$$\bar{T} = \frac{S^2 + \bar{r}}{\bar{r}} \tag{6}$$

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

In addition to reporting overall averages across 20 replications, a graphical analysis of standard error and RMSE conditioned on theta was conducted. The condition with the best efficiency, precision, and security indicators was selected, and its precision was compared with the linear test of Enem 2020.

Comparison with Enem 2020

With the response bank, we selected the responses to the items from Enem 2020 and calculated participants' scores with EAP. Then, we compared the indicators of precision of the linear test with the indicators of the selected CAT from the previous step.

Results

We present results regarding MT test, and graphs along with commands for all four areas are available at http://github.com/alexandrejaloto/PR_CAT.

Eficiency of the CATs

Table 3 presents the efficiency indicators for the eight variable-length CAT conditions. The MFI method showed smaller average test lengths (31.2) compared to PR2 (40.5) and PR3 (44.2) with the SE30 stopping rule. All three methods outperformed random, with differences up to 23 items (MIF). The average value of medians for the MFI method was 20, indicating that 20 items were sufficient to estimate the theta for 50% of the sample. Using progressive-restricted method, this value ranged from 40.1 (PR2) to 46.1 (PR3). There were no major differences between the methods in terms of the average values of the minimum length and maximum length of the tests in the four selection methods. ER015 rule resulted in smaller test lengths, with negligible

differences between selection methods.

Table 3. Average of the minimum, maximum, mean, and median values of items administered in the simulations

Stopping rule	Selection method	Min	Max	NIA	MIA
SE30	RAN	15.4	60.0	54.6	60.0
	MFI	15.0	60.0	31.2	20.0
	PR2	15.0	60.0	40.5	40.1
	PR3	15.0	60.0	44.2	46.1
ER015	RAN	15.0	20.8	15.3	15.0
	MFI	15.0	19.5	15.2	15.0
	PR2	15.0	23.5	15.7	15.0
	PR3	15.0	22.5	15.5	15.0

Note. 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 4 shows precision indicators for the 16 CAT conditions. MFI method exhibited the best precision values, while the random method showed the worst. Differences between MFI and PR methods were negligible, and also between acceleration parameters. Conditions with the ER015 stopping criterion had lower precision compared to SE30.

Table 4. Average of standard error, correlation, bias, and root mean squared error of the replications

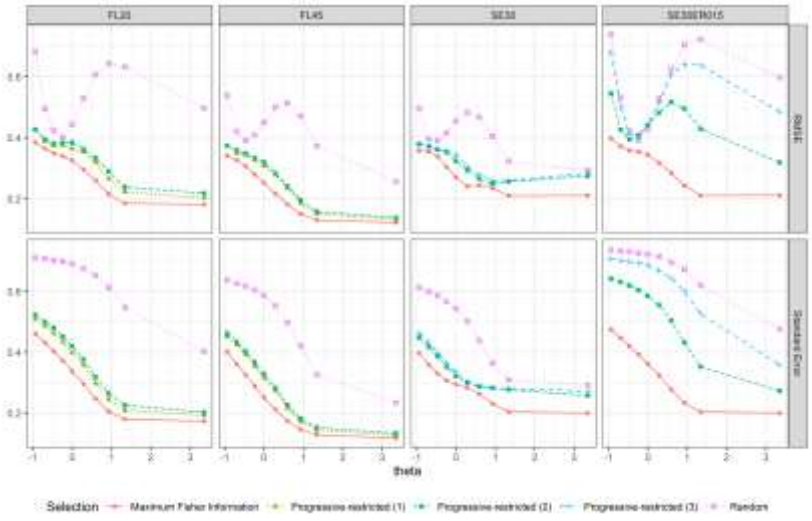
Selection method	SE	COR	Bias	RMSE	SE	COR	Bias	RMSE
Fixed-length CAT	FL45				FL20			
RAN	0.510	0.880	0.078	0.439	0.639	0.805	0.088	0.544
MFI	0.241	0.965	0.035	0.243	0.311	0.947	0.046	0.299
PR1	0.288	0.954	0.042	0.281	0.360	0.933	0.053	0.337
PR2	0.297	0.952	0.045	0.287	0.376	0.928	0.058	0.347
Variable-length CAT	SE30				ER015			
RAN	0.481	0.894	0.076	0.418	0.682	0.772	0.095	0.583
MFI	0.287	0.952	0.034	0.282	0.334	0.941	0.048	0.316
PR2	0.333	0.940	0.043	0.317	0.519	0.874	0.077	0.450
PR3	0.342	0.937	0.048	0.323	0.628	0.808	0.088	0.541

Note. 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 $k = 2$; PR3 = progressive-restricted with $k = 3$; FL20 = fixed-length (20 items); FL45 = fixed-length (45 items); 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 3 illustrates RMSE and standard error across the Enem scale. MFI, PR1, PR2, and PR3 showed significant gains in precision compared to the random method. Precision with PR as exposure control was adequate in fixed-length CATs and with the SE30 stopping rule, and with these rules there were no major difference between different acceleration parameters.

Figure 3. RMSE and 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 5 summarizes security indicators. In general, the random method provided the most secure test administrations, while the MFI method was the least secure. The inclusion of the PR exposure control method significantly increased test security, with higher acceleration parameter values resulting in greater security.

In fixed-length CATs with random, PR1, or PR2 methods, all items were presented at least once, irrespective of test size. However, the MFI method had non-administered items, indicating underutilization of the item bank. Exposure control methods capped maximum exposure rates at 0.30, and reduced item overlap rates.

For variable-length CATs under random or PR conditions, all items were presented at least once, regardless of the stopping rule. Similar to fixed-length CATs, the MFI method underutilized the item bank the most, having unadministered items. The inclusion of exposure control significantly reduced the proportion of non-administered and overexposed items.

Table 5. Mean of the minimum and maximum exposure rates and item overlap rate

Selection method	r_{min}	r_{max}	O	r_{min}	r_{max}	O
Fixed-length CAT		FL45			FL20	
RAN	0.035	0.098	0.059	0.014	0.038	0.026
MFI	0.000	1.000	0.432	0.000	1.000	0.404
PR1	0.005	0.300	0.181	0.001	0.300	0.157
PR2	0.011	0.300	0.149	0.003	0.300	0.125
Variable-length CAT		SE30			ER015	
RAN	0.041	0.131	0.072	0.011	0.029	0.020
MFI	0.000	1.000	0.406	0.000	1.000	0.406
PR2	0.007	0.300	0.167	0.006	0.127	0.033
PR3	0.014	0.300	0.139	0.009	0.045	0.021

Note. 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 $k = 2$; PR3 = progressive-restricted with $k = 3$; FL20 = fixed-length (20 items); FL45 = fixed-length (45 items); SE30 = standard error of 0.30; ER015 = standard error of 0.30 or error reduction of 0.015.

Table 6 details the mean percentage of items per exposure rate interval. The MFI method consistently had higher proportions of items not administered and items overexposed. Exposure control methods effectively reduced these proportions.

Table 6. Overall average percentages of items for each exposure rate interval in fixed-length CATs

Exposure	RAN	MFI	PR1	PR2	RAN	MFI	PR2	PR3
Fixed-length CAT	FL20				FL45			
0	0.0	82.6	0.0	0.0	0.0	62.4	0.0	0.0
(0;0.02]	0.0	5.8	81.6	80.4	0.0	13.6	55.2	18.6
(0.02;0.05]	100.0	2.7	7.4	9.2	25.8	5.1	19.8	57.8
(0.05;0.1]	0.0	2.0	3.5	4.4	74.2	5.2	8.8	10.4
(0.1;0.15]	0.0	1.5	2.1	1.5	0.0	1.8	3.3	3.2
(0.15;0.2]	0.0	1.3	1.5	1.4	0.0	1.6	2.3	2.1
(0.2;0.25]	0.0	0.6	0.8	1.3	0.0	1.8	2.3	1.5
(0.25;0.3]	0.0	0.3	2.5	1.6	0.0	1.3	5.7	4.4
(0.3;0.4]	0.0	0.9	0.5	0.1	0.0	2.0	2.7	2.0
(0.4;1]	0.0	2.4	0.0	0.0	0.0	5.3	0.0	0.0
Variable-length CAT	SE30				ER015			
0	0.0	78.5	0.0	0.0	0.0	87.4	0.0	0.0
(0;0.02]	0.0	6.7	52.4	1.9	69.4	4.0	76.0	70.6
(0.02;0.05]	0.4	2.4	28.2	78.7	30.6	2.0	17.6	29.4
(0.05;0.1]	96.8	1.5	5.9	6.8	0.0	1.0	5.6	0.0
(0.1;0.15]	2.8	1.1	2.7	3.3	0.0	1.3	0.9	0.0
(0.15;0.2]	0.0	0.9	2.4	2.0	0.0	1.1	0.0	0.0
(0.2;0.25]	0.0	1.6	1.4	1.9	0.0	0.3	0.0	0.0
(0.25;0.3]	0.0	2.0	5.7	4.5	0.0	0.5	0.0	0.0
(0.3;0.4]	0.0	1.8	1.4	0.9	0.0	1.0	0.0	0.0
(0.4;1]	0.0	3.4	0.0	0.0	0.0	1.4	0.0	0.0

Note. RAN = random; MFI = Maximum Fisher Information; PR1 = progressive-restricted with $k = 1$; PR2 = progressive-restricted with $k = 2$; PR3 = progressive-restricted with $k = 3$; FL20 = fixed-length (20 items); FL45 = fixed-length (45 items); SE30 = standard error of 0.30; ER015 = standard error of 0.30 or error reduction of 0.015.

Comparison between CAT and linear

The most satisfactory combination of CAT was the fixed length of 20 items with the PR2 method (PR2FL20). Although in terms of efficiency the performance of CATs with error reduction was higher than the fixed-length conditions, this decrease in test length was in no more than five items. Additionally, 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 7 presents the precision indicators of the linear test simulation and again of the PR2FL20 condition for ease of comparison. The precision obtained in the linear test was satisfactory, but they were better in the CAT.

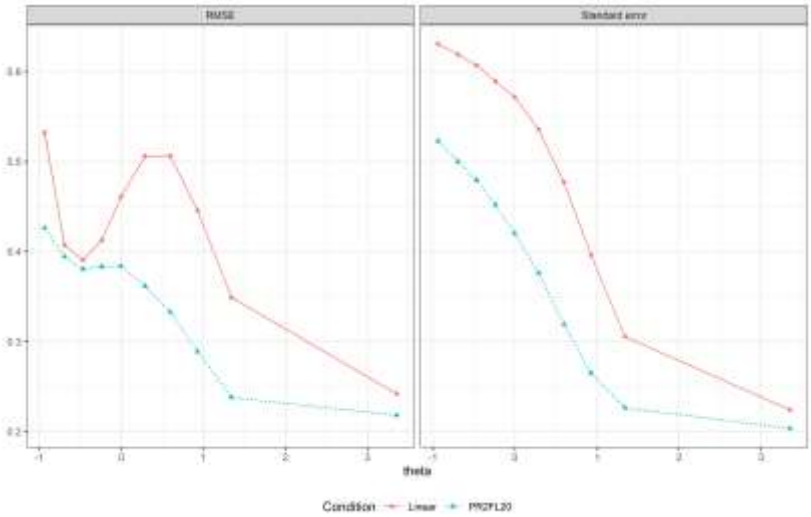
Table 7. Mean of standard error, correlation, bias, and root mean square error of the replications of simulations with linear application and PR2FL20 condition of the CAT.

Linear				PR2FL20			
Standard Error	Correlation	Bias	RMSE	Standard Error	Correlation	Bias	RMSE
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 4 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 precision compared to the linear application.

Figure 4. 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.