

Publication status: Preprint has been published in a journal as an article
DOI of the published article: <https://doi.org/10.18222/ae.v35.10142>

Next-generation Enem assessment with fewer items and high reliability using CAT

Alexandre Jaloto, Ricardo Primi

<https://doi.org/10.1590/SciELOPreprints.5339>

Submitted on: 2022-12-27

Posted on: 2023-01-03 (version 1)
(YYYY-MM-DD)

Next-generation Enem assessment with fewer items and high reliability using CAT

Alexandre Jaloto

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep),
Brasília, Brasil

Departamento de Psicologia, Universidade São Francisco, Campinas, Brasil

ORCID: <https://orcid.org/0000-0002-5291-1768>

Ricardo Primi

Departamento de Psicologia, Universidade São Francisco, Campinas, Brasil

ORCID: <https://orcid.org/0000-0003-4227-6745>

Abstract

The Brazilian High School Exam (Enem) results are used to enter higher education and obtain scholarships and study financing. The Enem comprises an essay and four 45-item tests that measure educational skills in different areas of knowledge. Therefore, this work aims to verify the possibility of reducing the number of items in Enem through a Computerized Adaptive Testing (CAT). We used tests from the 2009 to 2019 editions of Enem. The items have been calibrated and positioned on four scales (one for each area). we simulated a CAT, which ended when the error was less than or equal to 0.30, or 45 items were applied. On average, the application ranged from 12.0 (Languages and Codes; LC) to 29.2 (Mathematics; MT) items. Results point to the potential of reducing the size of Enem to 20 items for a proportion that varies from 39.8% (MT) to 94.8% (LC).

Keywords: student assessment; psychometrics; item response theory; admission to higher education

Introduction

The purpose of this study is to determine whether the Brazilian High School Examination (also known as Enem in Portuguese) can be simplified by using Computerized Adaptive Testing (CAT) without compromising reliability. Since 2009, Enem, which comprises an essay and four tests with 45 multiple-choice items, has been used as a single selection process for admission to several higher education institutions. Considering that the cutoff scores for each course vary widely, the reliability needs to be adequate within a range of about seven standard deviations. This study combines a reflection on the technical feasibility of implementing the CAT in the Enem in view of the impact the position of an item has on educational test performance (Debeer & Janssen, 2013; Domingue et al., 2020; Wu et al., 2016) with a consideration of the potential of CAT to improve the logistics of applying Enem. It also seeks to advance knowledge about CAT in educational tests by conducting a simulation with banks consisting of 674 to 839 items.

Enem

Enem was established in 1998 by the National Institute for Educational Studies and Researches Anísio Teixeira (in Portuguese, Inep), a Ministry of Education agency. Among its goals was to provide a framework for self-assessment based on performance indicators and the development domain of skills and abilities inherent in the cognitive and social development phase at the end of basic education (Inep, 2009).

Enem underwent changes in 2009 that made it closer to becoming the only entrance test for undergraduate study at Brazilian federal higher education institutions. Due to the Unified Selection System (in Portuguese, SiSU), which was established by the Ministry of Education, institutions select candidates based on their test scores. Moreover, participation in Enem has become a requirement to compete for public scholarships or funding to attend a private college or university.

The exam contains an essay and four 45-item tests, each of which combines skills from four different areas of knowledge, namely: Human Sciences (HS); Natural Sciences (NS); Languages and Codes (LC); and Mathematics (MT). Using Item Response Theory (IRT), each test is applied in a linear format and produces a one-dimensional measure for each area (Inep, 2012). In order to

select candidates for entry into their courses, educational institutions may use arithmetic or a weighted average of the four tests.

Implications of using Enem for selection

There are several challenges associated with the use of the Enem for selection in terms of reliability and validity. First, it is essential to demonstrate the validity of the test for predicting performance in higher education. Few studies have explored this aspect (see, for example, Ferreira-Rodrigues, 2015). A second issue is the assessment of the equivalence of essay scores, since judges differ in their levels of leniency and severity (Primi et al., 2019), which is not taken into account in the current version of the test. Finally, measurement error is a concern when using the test with candidates with very different proficiency levels. We do not know how large the error at each scale level is, with some courses having cutoffs at lower levels of the scale and others at higher levels.

Inep transforms and standardizes IRT scores on ENEM to have a mean of $M=500$ and standard deviation of $SD=100$, using as reference statistics of the 2009 edition (Inep, 2012). Considering the wide variation in cutoff scores, items must cover a broad spectrum of difficulty to obtain sufficient reliable scores for selection decisions. As an example, in the 2020 edition of SiSU, which incorporated the Enem 2019, the lowest average enrollment in the Social Sciences course at the Federal University of Rio de Janeiro for vacancies reserved for black students with disabilities and low-income public schools was 394.10. On the other hand, the lowest average to enter the Medicine course at the same institution was 790.98. Cutoff scores ranged from 227.78 (Aquaculture Engineering) to 928.30 (Medicine). In addition, to qualify for public scholarships and funding, a minimum average of 450 is required. Based on these cutoff scores, it is evident that the tests require adequate reliability over a range of more than seven standard deviation units depending on the Enem edition. This information and other information about the history of Enem and SiSU editions is available at www.gov.br/inep and www.sisu.mec.gov.br. Inep attempts to achieve Enem's objectives through tests with a fixed number of 45 items. Clearly, it is practically impossible to achieve the same level of reliability with 45 items over a range of seven standard deviations. CAT may be one of the few ways of achieving this objective.

Additionally, fatigue is also a concern. We know that in high-stakes large-scale educational achievement tests, a question's position can affect its properties of difficulty and discrimination (Domingue et al., 2020). As an example, an item placed in the final section of a test will have a lower hit rate than if it is presented to students at the beginning of the test. Similarly, participant

performance declines in the final items of the test (Debeer & Janssen, 2013; Wu et al., 2016). Also, in Enem 2016 the item position in MT was associated with the performance in the item (Barichello et al., 2022). These findings suggest that fatigue may interfere with student performance.

When there are individual differences in performance and these differences do not relate to the construct measured by Enem, this can result in an equity issue for the test. Students with similar levels of proficiency may score differently as a result of fatigue rather than the main competency. The application of a CAT could mitigate the problem of the size of a test and, consequently, reduce the potential effects of fatigue. The promise of CAT design is to maintain the reliability of a test with fewer items, or even enhance it (Veldkamp & Matteucci, 2013; Weiss, 2011).

Next-generation Enem assessment using CAT

In the IRT framework, the error is inversely related to the amount of information in the region of the estimated measurement level (Ayala, 2009). Generally, a question is most informative for people whose abilities are comparable to its difficulty. Therefore, too difficult or too easy items have little impact on the reliability of people with lower or higher abilities, respectively. However, due to the nature of Enem, which must be able to measure a large range accurately, participants with high ability levels are required to complete easy items, even though the information they add is minimal. Moreover, participants with low ability levels will be required to respond to questions with a very high level of difficulty. Alternatively, CAT selects items for administration based on the participant's provisional ability estimates computed from previous answers, so items that are too far from the participant's level of ability will be avoided.

By following the procedure, subjects will answer a large proportion of items that are useful to inform their level of proficiency and not waste their time answering uninformative items. This could solve a problem with the Enem MT test, which is off-target for the people taking it. This test is much more difficult than the average level of proficiency of most examinees taking Enem. As a result, the reliability of most scores is very low. Thus, it is difficult to differentiate between low-ability students.

Studies have confirmed that CATs can be used to reduce the size of educational tests. Kalender & Berberoglu (2017) simulated an admission test for higher education in Turkey consisting of 17 items on average. Originally, this test had 45 items. Spennassato et al. (2016) simulated the application of the 45 MT items of Enem 2012 in the format of a CAT and concluded that a 33-item extension of the CAT produced comparable results. Mizumoto et al. (2019)

demonstrated that English language vocabulary tests consisting of 115, 73, and 56 items could be reduced to tests consisting of 10, 15, and 20 items, respectively. Considering these are high-stakes tests, the item banks in these studies contained relatively few items for a CAT. In our study, we used item banks that ranged from 674 to 839 items.

Furthermore, in addition to reducing test size and measurement error, CAT can also significantly contribute to practical features of such a large-scale test. For example, one can pre-test items during the regular administration, reduce logistics complexity (no need to transport physical material), reduce chances of fraudulent disclosure of test content, and provide immediate feedback to participants.

The present study

The size of the test, which is one of the points mentioned above, was chosen as a starting point for exploring and reflecting on the feasibility of implementing CAT in Enem. So, the purpose of this research is to investigate whether it is possible to reduce the number of items in Enem through the use of a CAT without compromising measurement reliability. Our research was divided into two studies. First, it is intended to determine the item parameters applied in Enem and equate them to a single scale since Inep did not disclose the item parameters by the time this research was carried. The second study aims to simulate CAT analysis using these items.

Study 1

In this study, we set out to create an Enem item bank in a single metric as a guide to later performance on the CAT in Study 2. Three-parameter IRT logistic model was used to calibrate the parameters of the items from different years. We then converted them into a single metric using student scores from Inep's official database.

Participants

This study uses secondary data, extracted between January and June 2020 from Enem microdata, available on the Inep portal (www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados). We excluded participants whose microdata displayed some inconsistencies, such as a vector of responses with 44 instead of 45 characters. Additionally, we excluded participants who left the 45 answers blank.

We drew randomly selected samples of 5,000 participants from each application of the editions of Enem from 2009 to 2019. The samples were stratified based on the total score of correct responses. We therefore ensured that participants with high scores would be drawn and that items of most difficulty would be calibrated accordingly. We selected 1,250 participants from the lower and upper strata (25th and 95th percentiles, respectively) and 2,500 from the middle stratum. In applications with fewer than 25,000 participants, the upper stratum had fewer than 1,250 subjects ($0.05 * 25,000 = 1,250$). So, in these cases, the sample was supplemented with participants from other strata, maintaining a ratio of 1:2 (lower:intermediate). Due to rounding, some samples comprised more than 5,000 people.

There are at least two applications in each edition of Enem. Typically, the first application has the largest number of participants. Whenever there were fewer than 5,000 participants, calibration was performed with the entire population, if greater than or equal to 1,000. This sample size is suitable for calibrating items with the IRT three-parameter model (Şahin & Anıl, 2017) therefore we excluded applications with less than 1,000 participants. The sample draw was carried out using the *strata* function of the *sampling* package (v2.8; Tillé & Matei, 2016) in the R programming environment (R Core Team, 2019). In Table 1, we summarize the sample and population sizes of each application of Enem, along with other information that will be discussed in the Results section of Study 1.

Table 1

Information about each application (participants, sample, items used and correlation)

Area	Year	Application	Population	Sample	Number of excluded items	Number of used items	Correlation	
Human Sciences	2009	1	2,552,781	5,000	0	45	0.996*	
	2009	2	1,660	1,660	0	45	0.977*	
	2010	1	3,369,211	5,000	0	45	0.997*	
	2010	2	3,904	3,904	0	45	0.993*	
	2011	1	3,981,762	5,000	0	45	0.995*	
	2011	2	10,963	5,001	0	45	0.991*	
	2012	1	4,217,478	5,000	0	45	0.995*	
	2013	1	5,198,617	5,000	0	45	0.994*	
	2014	1	6,159,992	5,000	0	45	0.970*	
	2015	1	5,747,279	5,000	0	45	0.987*	
	2016	1	5,836,551	5,000	0	45	0.992*	
	2016	3	159,440	5,000	0	45	0.990*	
	2017	1	4,689,506	5,000	0	45	0.987*	
	2017	2	1,039	1,039	0	45	0.989*	
	2018	1	4,136,361	5,000	0	45	0.983*	
	2018	2	1,134	1,134	0	45	0.995*	
	2019	1	3,914,432	5,000	0	45	0.987*	
					Total	0	765	

Area	Year	Application	Population	Sample	Number of excluded items	Number of used items	Correlation	
Natural Sciences	2009	1	2,554,741	5,000	0	45	0.997*	
	2009	2	1,668	1,668	0	45	0.964*	
	2010	1	3,366,011	5,000	0	45	0.996*	
	2010	2	3,890	3,890	0	45	0.989*	
	2011	1	3,980,082	5,000	0	45	0.998*	
	2011	2	10,953	5,000	0	45	0.974*	
	2012	1	4,216,367	5,000	0	45	0.989*	
	2013	1	5,197,365	5,000	0	45	0.986*	
	2014	1	6,158,548	5,000	0	45	0.983*	
	2015	1	5,746,263	5,000	0	45	0.984*	
	2016	1	5,835,361	5,000	0	45	0.984*	
	2016	3	159,419	5,000	1	44	0.988*	
	2017	1	4,433,922	5,000	0	45	0.972*	
	2018	1	3,901,771	5,000	0	45	0.987*	
	2019	1	3,707,205	5,000	0	45	0.987*	
					Total	1	674	
	Languages and Codes	2009	1	2,434,642	5,000	1	44	0.999*
		2009	2	1,554	1,554	0	45	0.977*
		2010	1	3,246,005	5,000	0	50	0.998*
2010		2	2,306	2,306	0	50	0.994*	
2011		1	3,866,703	5,000	0	50	0.997*	
2012		1	4,090,691	5,000	0	50	0.982*	
2013		1	5,022,660	5,000	0	50	0.994*	
2014		1	5,971,721	5,000	0	50	0.996*	
2015		1	5,615,054	5,000	0	50	0.994*	
2015		2	1,038	1,038	0	50	0.995*	
2016		1	5,685,125	5,000	0	50	0.988*	
2016		3	155,250	5,000	0	50	0.986*	
2017		1	4,693,808	5,000	0	50	0.993*	
2017		2	1,039	1,039	0	50	0.996*	
2018		1	4,140,393	5,000	0	50	0.989*	
2018		2	1,134	1,134	0	50	0.990*	
2019		1	3,917,238	5,000	0	50	0.985*	
					Total	1	839	
Mathematics		2009	1	2,433,932	5,000	0	45	0.984*
	2009	2	1,549	1,549	0	45	0.973*	
	2010	1	3,244,895	5,000	0	45	0.997*	
	2010	2	2,301	2,301	0	45	0.981*	
	2011	1	3,865,301	5,000	0	45	0.995*	
	2011	2	10,364	5,001	0	45	0.979*	
	2012	1	4,088,847	5,000	0	45	0.997*	
	2013	1	5,020,489	5,000	0	45	0.990*	
	2014	1	5,968,542	5,000	0	45	0.986*	
	2015	1	5,612,869	5,000	0	45	0.988*	
	2015	2	1,038	1,038	0	45	0.984*	
	2016	1	5,683,429	5,000	0	45	0.992*	
	2016	3	155,197	5,000	0	45	0.994*	
	2017	1	4,433,702	5,000	0	45	0.987*	
	2018	1	3,901,617	5,000	1	44	0.985*	
	2019	1	3,707,065	5,000	0	45	0.987*	
					Total	1	719	

Note: Population size, the sample size used for calibration, number of items excluded, total items, and correlation between the reestimated and the original (official) score of the population in each application of each area of knowledge. * $p < 0.001$

Measures

Each of the four Enem tests contains 45 multiple-choice items with five options, of which only one is correct. A three-parameter logistic model of IRT is used to estimate scores, which are positioned on a scale with a mean of 500 and a standard deviation of 100. The scale's reference is the 2009 regular graduates of the public schools (Inep, 2012). Since 2010, the LC test has contained five foreign language items, and participants can select either English or Spanish as their foreign language of choice. Accordingly, each test in this area contains 50 questions. The other three tests (HS, NS, and MT) have 45 items each. Occasionally, Inep may exclude an item due to pedagogical reasons (for example, two correct answers). In this study, we used all of the tests used in the 2009-2019 editions of Enem that did not undergo adaptation (for instance, those for people with impaired vision). Additionally, we excluded tests with fewer than 1,000 participants, those which did not have microdata, and the second application of LC of Enem 2011 because the information concerning its key is inconsistent. Table 1 presents the number of items used and excluded for each area.

Data analysis

The calibration was carried out with the *mirt* package (v1.33.2; Chalmers, 2012). For discrimination, we used log-normal prior distributions (mean 0 and a standard deviation of 0.5, which ensures positive values), and for guessing, we used beta distributions with parameters 7 and 28 (this was centered around 0.2, which is suitable for items with five alternatives). We reestimated all participant scores for each application through the *mirt* package, with the *expected a-posteriori* (EAP) method. The calibration adequacy was evaluated by correlating the reestimated and official scores (published by Inep). According to the assumption of invariance of the IRT parameters, we expected these correlation values to be close to 1, and the only difference will be of scale center and scale (intercept and slope).

As mentioned, Inep did not disclose the item parameters, and the only information we had about the scale of a test was the official scores of the participants, which have been equated throughout the years. We used the sigma-mean method to place the items in the same Inep metric

(Hambleton et al., 1991). The official scores of each sample (Inep metric) were transformed to a mean of 0 and a standard deviation of 1. We also recalculated student scores using *mirt* in metrics 0 and 1 (*mirt* scores) in this study. Therefore, we had two theta values for each student (official and *mirt*), which should be identical. Yet they were not, as it is unlikely that the ability distribution of the selected sample had an exact mean of zero and a standard deviation of one. However, *mirt* assumes these values when calibrating/estimating for metric identification. Imagine that Y_i represents the official score (metric in which we want to calibrate the items) and X_i represents the *mirt* score for subject i . We can express the equality of these two scores as follows:

$$\frac{Y_i - \bar{Y}}{DP_y} = \frac{X_i - \bar{X}}{DP_x} \quad (\text{E1})$$

Where \bar{Y} and DP_y represent the mean and standard deviation of the official scores, and \bar{X} and DP_x represent the mean and standard deviation of the *mirt* scores. In this equation, we consider that the standardized scores of the two metrics are equal (or should be, as said), as they come from the same sample. If we isolate Y_i , we will have (Muñiz, 1997):

$$Y_i = \frac{DP_y}{DP_x} X_i + \left[\bar{Y} - \frac{DP_y}{DP_x} \bar{X} \right] \quad (\text{E2})$$

The constants k and d

$$k = \frac{DP_y}{DP_x} \quad (\text{E3})$$

$$d = \bar{Y} - k\bar{X} \quad (\text{E4})$$

extracted from this equation represent the equating constants (scale and origin, respectively) to transform the subjects' scores and the item parameters obtained via *mirt* to the Inep metric. As the subjects' scores and the item parameters are in the same metric, we can use these constants to transform the parameters b and a of item j from the *mirt* metric to the official Inep metric, using the following linear equations:

$$b_{j \text{ inep}} = kb_{j \text{ mirt}} + d \quad (\text{E5})$$

$$a_{j \text{ inep}} = \frac{a_{j \text{ mirt}}}{k} \quad (\text{E6})$$

The items in the same area are then all positioned on the official Inep metric following these transformations. Our results are therefore comparable to the official ones, and we can estimate a simulation participant's score by using all items in a particular area. Therefore, it is as if we had a test consisting of hundreds of items for each of the knowledge areas in Enem. The codes used in this work are available at <http://github.com/alexandrejaloto/CATENem.MFI>.

Results

Correlations between reestimated and official scores were statistically significant ($p < 0.001$) and high. The smallest was 0.96, observed in NS (2nd application of 2009). High magnitude correlations show that variations in reestimated scores explain variations in official scores. This indicates that the calibration in this study produced item parameters comparable to the official ones. One might question why these values do not reach one and differ only in scale and origin. There may be several reasons for this, including the issue of estimating a guessing parameter. According to our observations, the most discrepancies between our and official results are observed for students with low ability. This can be attributed to differences between guessing parameters. A recalibration may show differences in guessing more than other parameters (Primi et al., 2018). Figure 1 shows the official scores as a function of the reestimated scores in the first application of HS 2014, the lowest correlation observed in a principal application.

Figure 1

Relationship between the official and reestimated scores in the first application of Enem 2014 (HS)

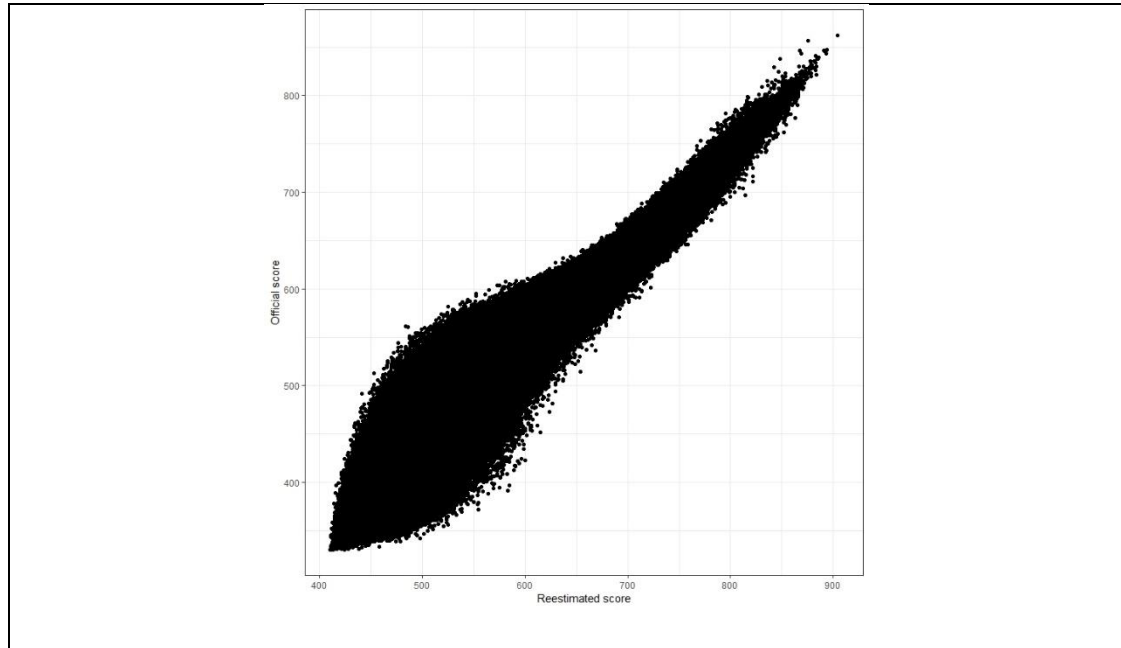


Table 1 brings all the correlation values between the reestimated and official scores. Once we obtained reasonable estimates of the parameters of the items applied in Enem, we moved on to Study 2. Due to high correlations, we can say that the simulation of Study 2 has results comparable with the results published by Inep.

Study 2

The purpose of this study was to simulate the application of CAT with items from all editions of Enem. To this end, we selected a random sample of participants from the 2019 edition to have an initial set of theta values. We then simulated responses from these subjects to all the items within the bank of items. Next, we conducted the CAT algorithm using the measurement error and/or the number of items applied as a stopping rule.

Participants

A simple random sample of participants from the 2019 edition of Enem was drawn for each area. The sample size was such as to guarantee a mean with a sampling error of 3 points on the Enem scale, which is equivalent to 0.03 standard deviation units. By adopting this procedure, we were able to generalize our results to the population of this edition of Enem, which potentially brings our simulation closer to expected situations for future editions with similar characteristics. Descriptive statistics of the participants of the 2019 edition and of the samples drawn from each area are

presented in Table 2.

The responses to the items were simulated using the `generate_pattern` function of the `mirtCAT` package (v1.10; Chalmers, 2016); input values were the official sample scores (we called them true scores) and item parameters. As a result, we obtained four response banks, one for each area of knowledge. Each response bank contained the number of lines corresponding to each area of knowledge sample size and the number of columns corresponding to the size of the item bank in each area. We produced a response bank as if each subject had answered all items in an area.

Table 2

Descriptive statistics of the participants of the 2019 edition of Enem and of their samples

Area	n	Mean	Standard Deviation	Range
Human Sciences				
Participants in 2019	3,917,245	508.0	80.1	315.9–835.1
Simulation sample	2,738	509.0	79.9	321.5–771.1
Natural Sciences				
Participants in 2019	3,709,827	477.9	75.9	327.9–860.9
Simulation sample	2,457	478.1	75.1	329.2–737.0
Languages and Codes				
Participants in 2019	3,920,058	520.9	62.5	322.0–801.7
Simulation sample	1,667	522.1	62.0	325.6–688.0
Mathematics				
Participants in 2019	3,709,686	523.2	108.8	359.0–985.5
Simulation sample	5,055	522.4	107.8	359.0–929.2

Measures

We used the four item banks obtained in the previous study. The HS bank contained 765 items, the NS, 674 items, the LC, 839 items, and the MT, 719 items.

Analysis

Data analysis for Study 2 consisted of the CAT simulation, which was performed using the `mirtCAT` package. Items were presented according to the maximum information criterion. The estimation method was the EAP. We called the scores estimated in CAT simulation as estimated scores. Simulation ended when the standard error of measurement reached a value of 0.30 or when 45 items were applied. This standard error value corresponds to a reliability of 0.91, since (Nicewander & Thomasson, 1999)

$$\sigma_{\hat{\theta}} = \sqrt{1 - \rho(\hat{\theta})} \tag{E7}$$

where $\rho(\hat{\theta})$ is the reliability for a given score and $\sigma_{\hat{\theta}}$ is the standard error of measurement.

By substituting the error value, we have

$$0.30 = \sqrt{1 - \rho(\hat{\theta})} \therefore \rho(\hat{\theta}) = 0.91 \tag{E8}$$

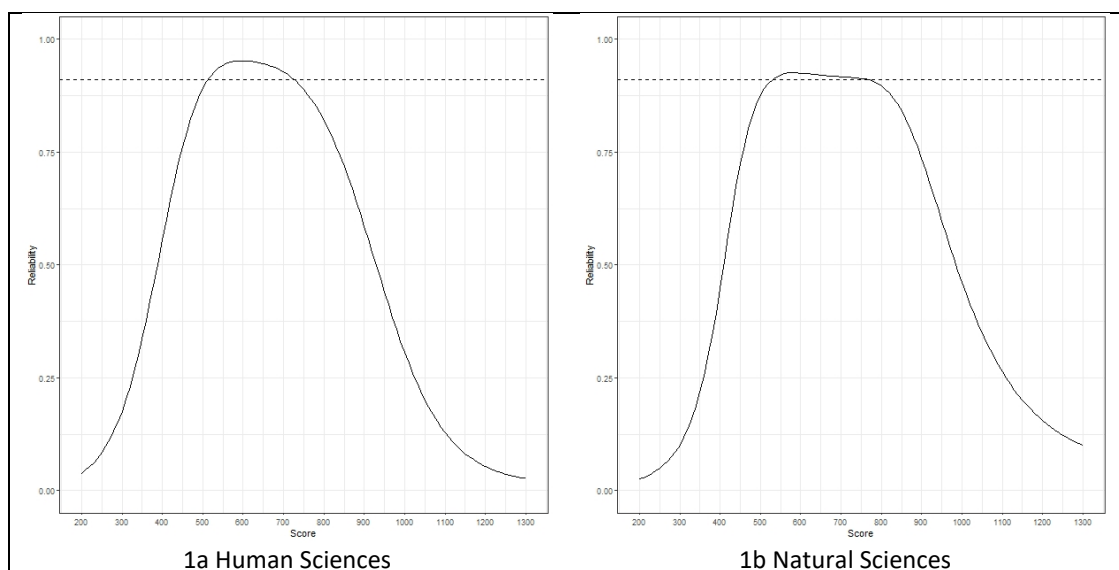
The criterion we used was more stringent than the highest marginal reliability found in Enem’s first application in 2019 ($HS = 0,78$; $NS = 0,74$; $LC = 0,84$; $MT = 0,53$). Marginal reliability reflects an average precision across the whole scale (Ayala, 2009). The reliability of 0.91 also exceeds the precision on most of the scales of these tests. The precision at each point of the scale $\rho(\theta)$ was calculated based on information $I(\theta)$ from the test for this point (Nicewander & Thomasson, 1999):

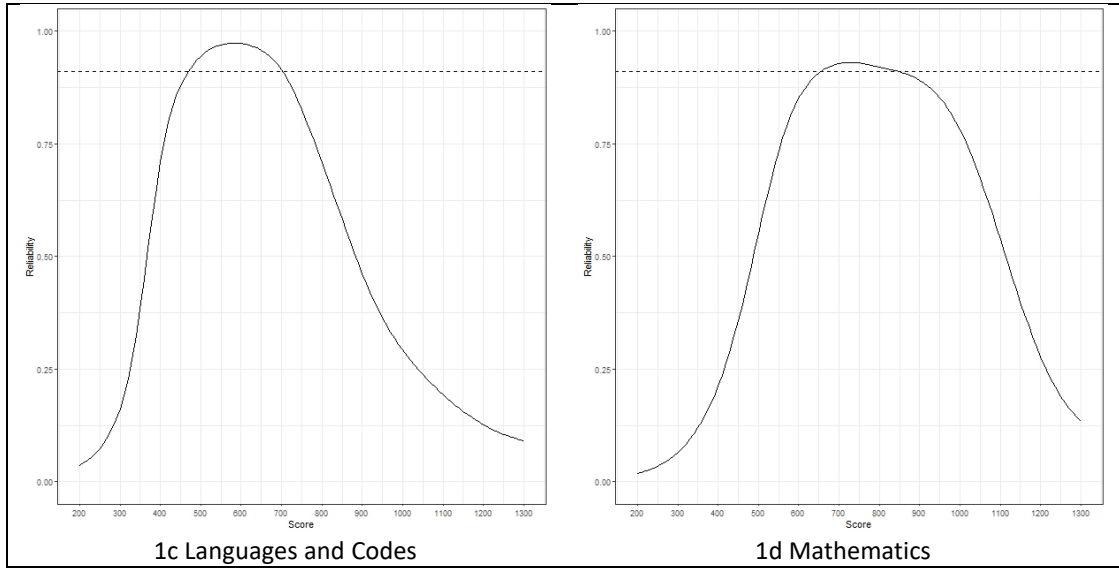
$$\rho(\theta) = \frac{I(\theta)}{I(\theta)+1} \tag{E9}$$

Figure 2 shows the reliability as a function of the score in the tests of the first application of Enem 2019.

Figure 2

Reliability of tests from the first application of Enem 2019





Note: the horizontal line indicates the reliability of 0.91.

Results

Across all test areas, we observed a reduction of at least 40% in test size for more than half of the participants and a maximum error of 0.30. The HS test was reduced to a maximum of 20 items for 71.7% of the participants. The same reduction in the NS, LC, and MT tests was possible for 60.4%, 94.8%, and 39.8% of the participants. With 15 items, it was possible to estimate the score of more than half of the participants in HS, NS, and LC with a maximum error of 0.30. Around 30 items were needed in MT. Additionally, 90.3% of participants in HS, 79.6% in NS, 99.2% in LC, and 65.6% in MT reduced at least one item on their test. On average, 18.4 items were administered in HS, 22.1 in NS, 12.0 in LC, and 29.2 in MT. Table 3 shows the percentage of participants in each area that underwent a maximum of 15, 20, 30, and 44 items and the average number of items applied in each test.

Table 3

Results of simulation (correlation, error, and number of applied items)

Area	Correlation	Average standard error	Highest standard error	Minimum number of items	At last 15 items	At last 20 items	At last 30 items	At last 44 items	Mean of Applied items	Total of items in bank
Human Sciences	0.937*	0.29	0.47	8	58.7%	71.7%	83.1%	90.3%	18.4	765
Natural Sciences	0.922*	0.31	0.53	8	52.3%	60.4%	70.9%	79.6%	22.1	674
Languages and Codes	0.904*	0.29	0.43	7	89.6%	94.8%	97.8%	99.2%	12	839
Mathematics	0.957*	0.32	0.56	11	25.8%	39.8%	54.0%	65.6%	29.2	719

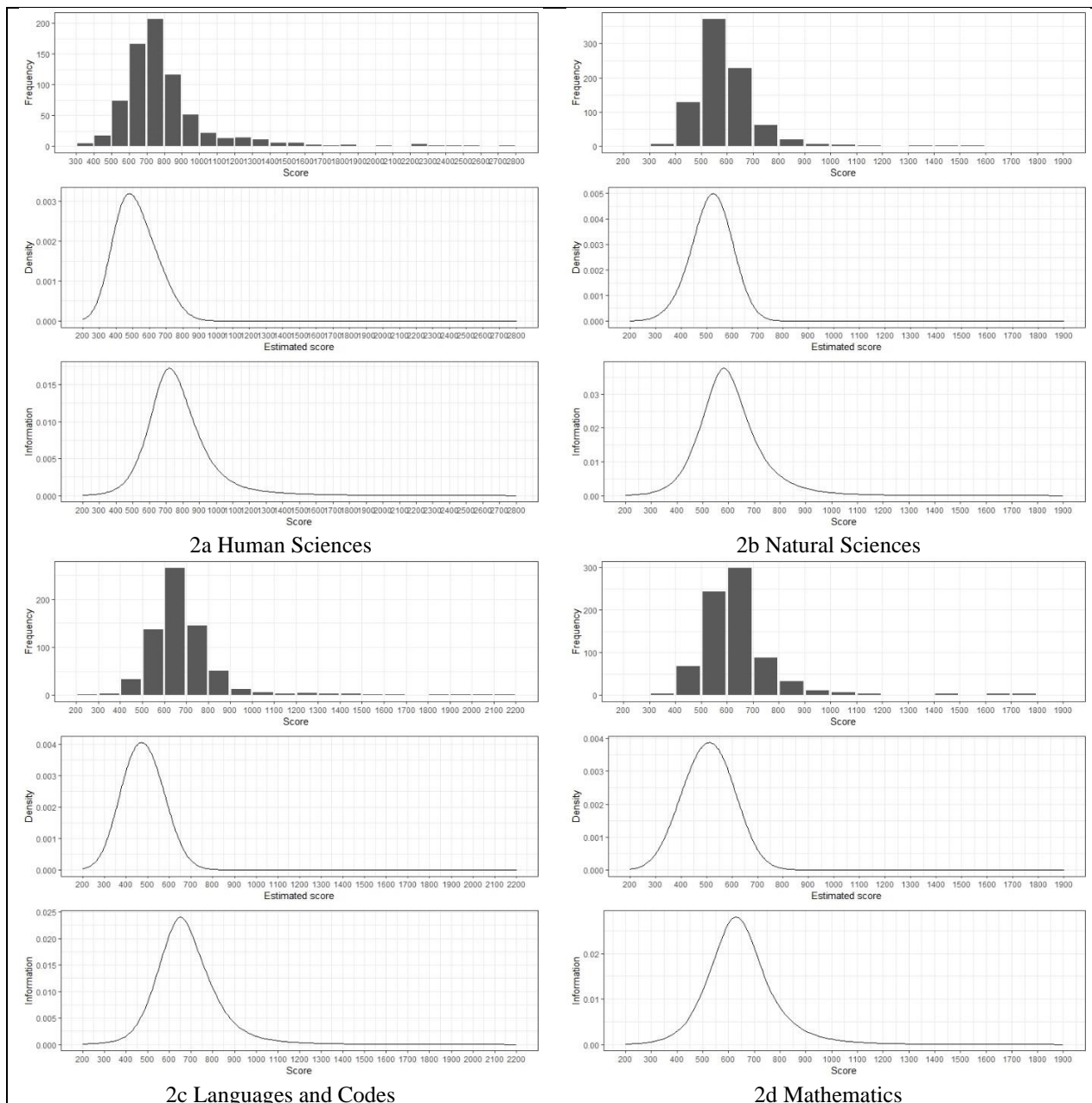
Note: Correlation between estimated scores and true scores, mean standard error of measurement, highest standard error, the lowest number of applied items, percentage of participants submitted to a maximum of 15, 20, 30, or 44 items, mean of applied items, and total quantity of items in the bank of each area. * $p < 0.001$

We also found cases where the error was greater than 0.30 even with 45 items applied. The average standard error was 0.29 in HS and LC, 0.31 in NS, and 0.32 in MT. The biggest error observed in HS was 0.47; in NS, 0.53; in LC, 0.43; and in MT, 0.56. Even though not all errors were below 0.30, the estimated scores showed a significant correlation with the true scores. Using an optimized method of selecting items to create a test form tailored to each subject produced results similar to those that would have been achieved if all items had been answered. Table 3 shows the values of errors and correlations, the minimum number of items applied in the simulation of each area, and the total number of items in each bank.

Figure 3 shows the number of items in the bank as a function of their location on the scale, the distribution of estimated score, and the item bank information curve. Considering the location of items on the scale, in LC, the region with the greatest number of items is close to the mean; in HS and NS, close to one standard deviation unit above the mean; and in MT, close to two standard deviation units above the mean. Among the four areas, these are the regions with the most information. Nevertheless, a great portion of the distribution of theta is in regions with limited information. This arises from our choice of theta distribution, which should mimic the distribution of real Enem data observed in 2019. It is evident that Enem's item database is off-target upward — it contains items of disproportionately greater difficulty. There are fewer items available that are capable of measuring students with low abilities in a precise manner.

Figure 3

Number of items, distribution of estimated scores, and item bank information curve for all areas



Note: In each quadrant, the upper figure shows the number of items as a function of its location in scale, the middle figure shows the distribution of estimated scores, and the lower figure shows the item bank information curve.

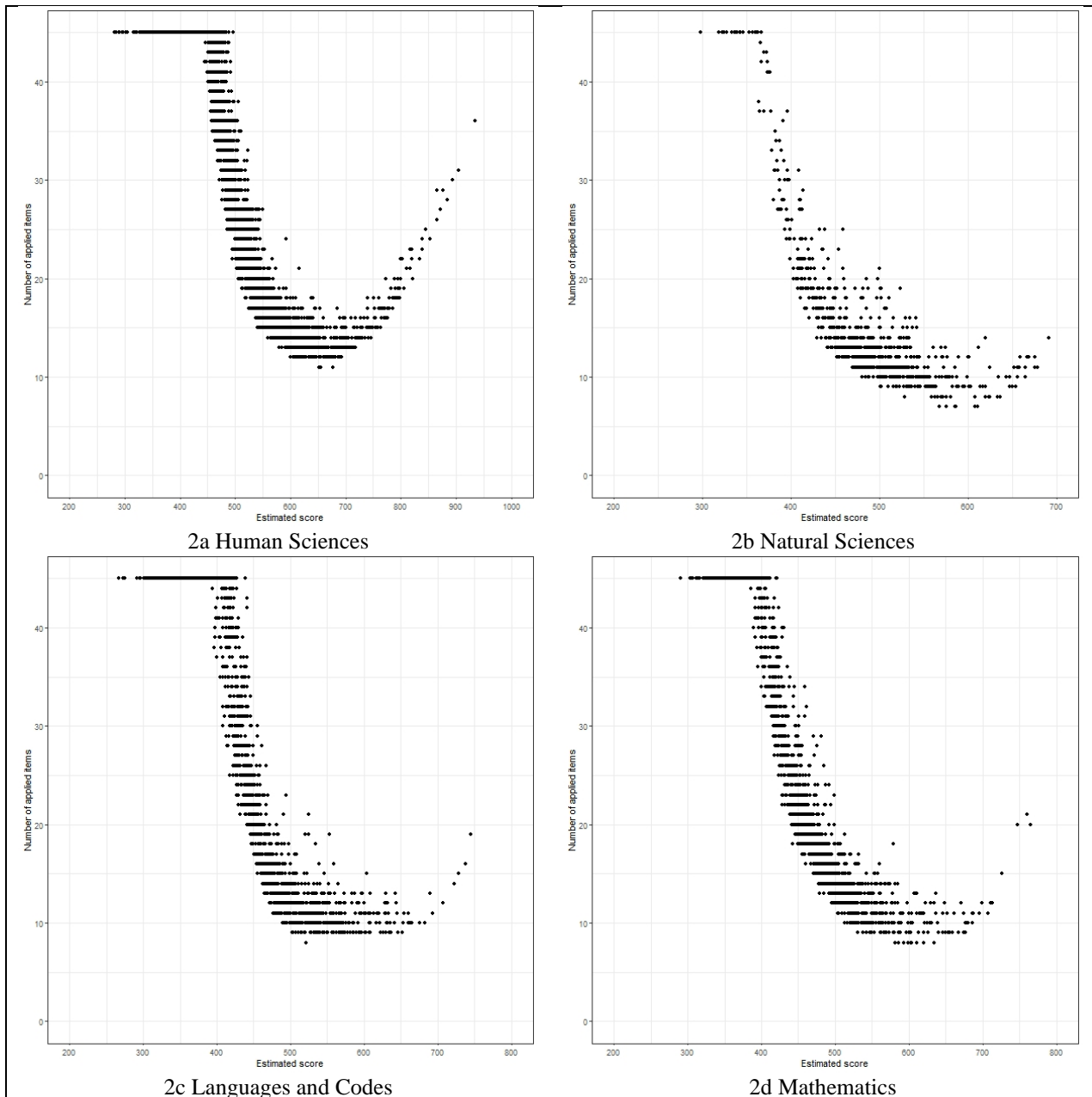
The CAT algorithm tended to give more items to participants located in regions with lower levels of information. In Figure 4, the number of items presented in the CAT is plotted against estimated score. It is evident that 20 items were sufficient for participants with theta in the range of 499.1—747.8 (in HS); 525.0—745.4 (in NS); 500.0—691.0 (in LC); and 615.3—798.4 (in MT).

In lower regions of the scale, CAT algorithm administered more items. For example, in HS, all participants with theta less than 385.1 were presented to 45 items; in NS, 393.6; in LC, 348.0;

and in MT, 445.1. The highest theta of a participant submitted to 45 items in HS was 421.0; in NS, 439.3; in LC, 366.7; and in MT, 495.8. The test size of all participants with scores higher than these was reduced by at least one item.

Figure 4

Number of items presented according to the estimated score in each test



Discussion

In this study, the goal was to determine whether it is possible to reduce the number of items in Enem by using a CAT without compromising reliability. Results show that, in the four knowledge

areas assessed in Enem, a reduction would be possible for a wide range of scales. The simulation ended with a maximum of 20 items in scale ranges that corresponded to 2.5 standard deviation units in HS; 2.2 standard deviation units in NS; 1.9 units in LC; and 1.8 units in MT. With a reliability of .91, 20 items could estimate 71.7% of participants' scores in HS; 60.4% in NS; 94.8% in LC; and 39.8% in MT. In all tests, the application did not reach 30 items for more than half of the participants.

The average size of the HS application was close to that of the simulation conducted by Kalender & Berberoglu (17; 2017), who used a 45-item bank of a Turkish higher education admission test with accuracy as the criterion of stop (0.30 error). They observed correlations between the CAT-based and original scores (linear test), ranging from 0.66 to 0.93. Our correlations all exceeded 0.90, which we expected, as we had a higher number of items in the bank.

We found that our simulation was more accurate than that of Spenassato et al. (2016), which simulated a CAT of 45 items of MT from Enem 2012. With a fixed application of 33 items, they observed a correlation of 0.99 between the CAT and the original score (linear test) and a mean error of 0.35. In the linear application, the authors observed errors of up to 0.84. In our simulation, we found a maximum error of 0.56 with an average of 0.32 in MT. In contrast to their study, our item limit was 45, but for 54.0% of the sample, 30 items were sufficient for an error of 0.30.

The article by Mizumoto et al. (2019) examined the possibility of reducing a vocabulary test of English which consists of three parts of 115, 73, and 56 items, for a fixed application of 20, 15, and 10 items, respectively. The proportion of samples with a measurement error of up to 0.33 ranged from 69.61% to 74.34%. HS (71.7%) with up to 20 items and NS (70.9%) with up to 30 items have similar proportions. Our proportion of LC applications with 15 items or less was higher (89.6%). In MT, 44 items were required to reach a similar percentage (65.6%).

It is important to highlight that our results indicate that the two tests with the most items (HS and LC) and more targeted items (i.e., items whose difficulties distribution matches theta distribution) provided the greatest reductions. This emphasizes the importance of a robust bank for the CAT to reduce the test size effectively. A noteworthy point is, however, that the variation in the LC sample and population was the smallest among the four groups. Consequently, it is not necessary to cover more extreme regions of the scale. We conclude that in cases where the population has a low variance, the item bank may be smaller than in cases where the population varies greatly. That is, the narrower the area where the population score is located, the fewer items one needs to estimate their score accurately. The smaller reduction in MT than in NS reinforces this

point. The first area had a larger bank, but the population variance was also greater. Future studies can verify the impact of the size of the variance on the need to increase the item bank.

In this study, the results were similar in all tests. Information curve peaks are located near the center of the scale or a little to the right, thus increasing the measurement accuracy of people residing in this range. As a result, the possibility of test reduction occurs mainly for participants whose theta are equal to or greater than the average (500). Therefore, our results show a limitation for reducing Enem in the entire range of the four scales.

It should be noted that the simulation bank had hundreds of items, and the algorithm selected those that contributed the most to the accuracy of the participant's score. However, for some subjects, even the 45 items (the same as that of the Enem tests) could not achieve an error of 0.30. Therefore, even if the subject had answered the linear test, the error of their score would have probably been greater than 0.30. The CAT failing to reduce the test size for these subjects is not due to a limitation of the CAT algorithm but rather due to the item bank limitation in covering lower levels of theta equally.

Prior studies have also found limitations to limiting testing to certain ranges of the scale. For instance, Kalender & Berberoglu (2017) reduced the test to 25 items for scale intervals ranging from 2.8 to 4.3 standard deviations. In our study, the reduction to 20 items occurred at intervals ranging between 1.8 and 2.5 standard deviations. Mizumoto et al. (2019) found that most scores with errors above 0.33 in their CAT simulation were above the scale mean. According to Spensato et al. (2016), the region below the mean had the highest average errors.

Given the small number of items (a maximum of 115) in the simulations described above, it is easier to understand the difficulties of achieving an adequate degree of precision. In our study, however, the banks contained at least 674 items. Nevertheless, we observed a similar pattern. In light of this, there is a possible limitation of large-scale educational tests, which results in excess of difficult items to discriminate high-ability students but do not have the same number of items that are easy and do a better differentiation of lower ability students. With respect to Enem specifically, the low precision in the lower region of the scale may affect the selection of courses with a low candidate-to-place ratio. Furthermore, it can also have a negative impact on the provision of public scholarships or funding. Consequently, this study reinforces the need for developing easier items for Enem.

Inep has developed guidelines for preparing and reviewing items (Inep, 2010, 2012). However, there are no objective guidelines on how to write items with predefined levels of

difficulty. In order to support the orientation of their objective elaboration, we recommend investigations into the factors associated with the items' difficulty. There has already been some evidence that aspects of a large-scale educational assessment item may be related to its difficulty (Kan & Bulut, 2015; Masri et al., 2017).

It is a limitation of this study that the item exposure was not controlled, which would be crucial in a real field administration scenario. Moreover, the study did not consider the representativeness of the content of knowledge areas (for example, curriculum topics and foreign language), as the only criteria for presenting the items was their information. Lastly, our study uses simulated responses that fit the models used. Therefore, our results might be biased towards an overly optimistic scenario. Considering this, we recommend that future studies fill these gaps and include other criteria for stopping and selecting items in simulations.

A possible CAT implementation agenda in Enem needs to include studies that identify how fatigue reduces the measurement accuracy. Several studies have found a relationship between item difficulty and its position on the test (Setzer et al., 2013; Ulitzsch et al., 2020), but there are no studies using tests with characteristics similar to those of Enem (i.e., high stakes and composed of four independent measures). Although we have verified the possibility of reducing the Enem in this study, we do not know if it is enough to minimize the effect of the fatigue on the answers.

Conclusion

Our study indicated that the Enem test could be reduced to 20 items for proportions ranging from 39.8% to 94.8% of participants by using a CAT. The reduction was most effective in domains where more items matched the distribution of proficiency skills in the population. By producing easier items in each of the four areas, the range of scale in which this reduction was possible can be expanded. We recommend that the guidelines for producing items include objective aspects of the item related to its difficulty. We hope that this work contributes to the improvement of large-scale educational assessments, particularly Enem. Moreover, we expect the selection process for admission to universities to become increasingly cheaper, more efficient, and fairer, as well as to include people whose right to access is currently denied.

Authors' contributions

Alexandre Jaloto designed the study, obtained the data, conducted the analysis and wrote most of

the manuscript. Ricardo Primi contributed to interpretation of the results and wrote the manuscript. All authors read and approved the final manuscript.

Conflicts of interest

Although Alexandre Jaloto works at the institute responsible for the exam, we have no conflicts of interest to disclose. The institute responsible for this assessment is governmental and has no profit ends. Alexandre Jaloto is a public employee and works at this institution directly with the psychometric analysis of Enem. Our research had no funding.

Disclosure statement

This text does not necessarily represent Inep's point of view. 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.

Acknowledgments

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

References

- Ayala, R. J. de. (2009). *The theory and practice of item response theory*. The Guilford Press
- Barichello, L., Guimarães, R. S., & Figueiredo Filho, D. B. (2022). A formatação da prova afeta o desempenho dos estudantes? Evidências do Enem (2016). *Educação e Pesquisa*, 48, e241713. [10.1590/s1678-4634202248241713por](https://doi.org/10.1590/s1678-4634202248241713por)
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. <http://doi.org/10.18637/jss.v048.i06>
- Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, 71(5), 1–39.

- Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, 50(2), 164–185. <https://doi.org/10.1111/jedm.12009>
- Domingue, B., Kanopka, K., Stenhaug, B., Sulik, M., Beverly, T., Brinkhuis, M. J. S., Circi, R., Faul, J., Liao, D., McCandliss, B., Obradovic, J., Piech, C., Porter, T., Soland, J., Weeks, J., Wise, S., & Yeatman, J. D. (2020). Speed accuracy tradeoff? Not so fast: Marginal changes in speed have inconsistent relationships with accuracy in real-world settings. *PsyArXiv*. <http://doi.org/10.31234/osf.io/kduv5>
- Ferreira-Rodrigues, C. F. (2015). *Estudos com o Enem a partir de uma abordagem psicométrica da inteligência* (Tese de Doutorado, Universidade São Francisco, Itatiba, SP). <https://www.usf.edu.br/galeria/getImage/427/2977366806369866.pdf>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2009). *Exame Nacional do Ensino Médio (ENEM): Textos teóricos e metodológicos*. MEC/INEP
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2010). *Guia de elaboração e revisão de itens*. MEC/INEP. http://download.inep.gov.br/outras_acoes/bni/guia/guia_elaboracao_revisao_itens_2012.pdf
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2012). *Entenda a sua nota no Enem: Guia do participante*. MEC/INEP. http://download.inep.gov.br/educacao_basica/enem/guia_participante/2013/guia_do_participante_notas.pdf
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2012). *Guia de elaboração de itens Província Brasil*. MEC/INEP.

http://download.inep.gov.br/educacao_basica/provinha_brasil/documentos/2012/guia_elaboracao_itens_provinha_brasil.pdf

- 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>
- Kan, A., & Bulut, O. (2015). Examining the language factor in mathematics assessments. *Journal of Education and Human Development*, 4(1), 133–146. <http://doi.org/10.15640/jehd.v4n1a13>
- Masri, Y. H. E., Ferrara, S., Foltz, P. W., & Baird, J.-A. (2017). Predicting item difficulty of science national curriculum tests: The case of key stage 2 assessments. *The Curriculum Journal*, 28(1), 59–82. <https://doi.org/10.1080/09585176.2016.1232201>
- 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](https://doi.org/10.1177/0265532217725776)
- Muñoz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Psicología Pirámide
- Nicewander, W. A., & Thomasson, G. L. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement*, 23(3), 239–247. <https://doi.org/10.1177/01466219922031356>
- Primi, R., Nakano, T. de C., & Wechsler, S. M. (2018). Using four-parameter item response theory to model human figure drawings. *Revista Avaliação Psicológica*. [10.15689/ap.2018.1704.7.07](https://doi.org/10.15689/ap.2018.1704.7.07)
- Primi, R., Silvia, P. J., Jauk, E., & Benedek, M. (2019). Applying many-facet Rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2), 176–186. <http://doi.org/10.1037/aca0000230>

- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, 17(1), 321–335. <http://doi.org/10.12738/estp.2017.1.0270>
- Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, 26(1), 34–49. <http://dx.doi.org/10.1080/08957347.2013.739453>
- 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>
- Tillé, Y., & Matei, A. (2016). *Sampling: Survey Sampling*. <https://CRAN.R-project.org/package=sampling>
- Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A multiprocess item response model for not-reached items due to time limits and quitting. *Educational and Psychological Measurement*, 80(3), 522–547. <https://doi.org/10.1177/0013164419878241>
- Veldkamp, B. P., & Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio: Avaliação e Políticas Públicas em Educação*, 21(78), 57–72. <https://doi.org/10.1590/S0104-40362013005000001>
- Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, 2(1), 1–27. <https://doi.org/10.2458/v2i1.12351>
- Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers Theory into practice*. Springer

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.