Recital
Revista de Educação,
Ciência e Tecnologia de Almenara/MG.

OS TESTES SITUACIONAIS DE INTELIGÊNCIA EMOCIONAL COMO UM TESTE ADAPTATIVO COMPUTADORIZADO

The Situational Tests of Emotional Intelligence as Computer-Adaptive Tests

Victor Vasconcelos de SOUZA
Universidade de Brasília
contato@dr.victorvs.com

Cristiane FAIAD
Universidade de Brasília
crisfaiad@gmail.com

DOI: https://doi.org/10.46636/recital.v5i3.478

Abstract
The present study aimed to assess whether an emotional understanding test (STEU) and an emotional management test (STEM) could benefit from being administered as computer-adaptive tests (CATs) without impacting the validity of the test scores. To this end, 11 item selection algorithms (ISAs) were benchmarked for their bias and efficiency. Two simulation studies were run using the same response patterns from the 688 participants used in the original validation study, the same 11 ISAs, but differed in their stopping rules (SRs). For the first study, one simulation was run for each ISA with the SR being standard error lower than 10^{-3} ($\Delta SE < 10^{-3}$), the most commonly used stopping rule criterion. For the second study, k simulations were run for each ISA, for each test, with the SR being a fixed number of items between 1 and k, where k was the total number of items of the relevant test (32 for the STEU and 30 for the STEM).

Results of the first simulation showed that testing with all ISAs resulted in accurate ability estimates, all of them having $r > .98$ between their estimates and the estimates calculated with the entire test. The first simulation also showed that, using the best performing ISA, 368 (53.5%) participants needed to answer at least one fewer item without loss of validity. The second study showed that the STEU stood to benefit the most, with the mean standard error (MSE) being minimized six items before the end of the test, though ISAs based on the Kullback–Leibler information performed worse. However, these ISAs also displayed slightly less bias, $r \approx .99$, than the Fisher information-based ones $r \approx .98$. For the STEM, no ISA minimized MSE levels before the end of the test, but up to six fewer items for the STEM and 15 fewer items for the STEU could be administered with a slightly higher tolerance of $\Delta SE < 10^{-2}$. These results indicate that the use of CAT methodology to administer these tests is viable, and EI testing stands to gain from using CAT tests. Future studies should test ISA performance with additional testing constraints.

Keywords: computer-adaptive testing, situational tests of emotional intelligence, emotional intelligence, item selection algorithms, item response theory.

1 INTRODUCTION

The Situational Test of Emotion Understanding (STEU) and the Situational Test of Emotion Management (STEM) are two ability EI tests initially described by MacCann and Roberts (2008). The tests measure two dimensions of the Mayer–Salovey–Caruso (MSC) EI theory (Mayer et al., 2012) and were designed partly in response to the fact that, despite the success of the MSC theory, almost all studies conducted within the ability EI paradigm have utilized the Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT; Mayer et al., 2003; Vieira-Santos et al., 2018). While this has granted benefits such as comparability across studies and cultures, it has also introduced several problems, which MacCann and Roberts (2008) sought to address. Namely, the difficulty in discriminating construct variance from test variance, and the employment of two scoring systems which are mutually contradictory. Despite further discussions in which the authors admitted the problems (Mayer; Caruso; Salovey, 2016), attempts to address the problems with ability EI testing, of which the MSCEIT is the preeminent test, are still an active topic of research (Fiori; Agnoli; Davis, 2023). In addition to the problem addressed by MacCann and Roberts (2008), one of the proposed solutions for improving the validity of emotional intelligence testing overall is the use of computer-adaptive testing (CAT; Mancini et al., 2022).

A recent study sought to make both the STEU and the STEM available for use in a Brazilian setting, and after the cross-cultural adaptation procedure, the final test forms yielded favorable validity evidence (Souza, 2023).

Given the novel advantages of computer adaptive testing (CAT) applied to psychological tests, such as increased validity and precision, immediate feedback and a reduction of the exposure effect, and in an effort to combine the contributions of the STEU and STEM tests and of CAT, the present study aimed to evaluate whether the Brazilian version of the STEU and the STEM can be efficiently administered using CAT algorithms. Furthermore, we aimed to find out which item selection algorithms (ISAs) would most benefit from such testing.

1.1 COMPUTER ADAPTIVE TESTING

Although increasingly common, CAT is still not a widespread technique in the field of educational and psychological testing and is confused with various other terms that involve the use of testing in computer, such as computer-based testing and computer-delivered testing. CAT itself is a form of computer-delivered testing that employs an item selection algorithm (ISA) to administer a test in a way that is adaptive to the user's responses (Luecht, 2016). CAT is contrasted with linear testing, also known as fixed-form testing, the traditional type of testing in which test items are presented in a predetermined order (Luecht; Sireci, 2011).

The fundamental principle of CAT is straightforward: the testing platform initially presents the examinee with a question of average difficulty. If the examinee answers correctly, the ISA selects an item with greater difficulty. Conversely, if the examinee answers incorrectly, the algorithm selects an easier item. As the examinee answers more items close to their true aptitude, measurement error decreases (Luecht; Sireci, 2011). The testing session will end when measurement error reaches a predetermined, low value.

One of the most important properties of an adaptive test is the algorithm employed to select the items that are presented to the examinee, the ISA. ISAs employ a criterion to enumerate which item should be presented to the examinee after they answer the previous item. The two most common sources of information for the measures that are typically used as criteria by ISAs are the Fisher information, or FI (Lord, 1980) and the expected posterior variance, or EPV (van der Linden, 1998).

The FI is the amount of information that a given item extracts for a given theta interval, represented by $I_j(\theta)$, which is the information function at item *j* for ability θ. It is given by the probability function multiplied by the square of the logarithm of the probability function of answer *x* given ability θ (Frieden; Gatenby, 2013). An ISA employing the Maximum Fisher Information (MI) criterion calculates the FI for each item and selects the items with the largest value. The formula for calculating the FI can be viewed in Equation 1.

$$I_j(\theta) = \int_{\mathbb{R}} (log\, p(x|\theta))^2 \cdot p(x|\theta) dx \quad (1)$$

The EPV (Bock; Mislevy, 1982) measures the opposite of the FI of the item, that is, it measures the uncertainty associated with the estimation that can be made after that item is administered.

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5 n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

This is why it mathematically equivalent to the inverse of the FI, as it can be seen in Equation 3. Alternatively, the EPV measure, at item *j,* can be calculated by the arg min of the difference between the expected value of the ability θ, given the responses to the items that have been administered, and the expected value of the ability θ, given the responses to the items that have been administered plus the next item. This can be seen in Equation 2.

$$EPV = \frac{1}{I_j(\theta)} \tag{2}$$

$$EPV = argmin\left(E(\theta|x_1, \dots, x_k) - E\left(\theta\middle|x_1, \dots, x_j\right)\right) \tag{3}$$

It is, therefore, a measure of how much uncertainty will have shrunk after administering item *j*. The Minimum EPV criterion (van der Linden, 1998), selects items that minimize it. It is, however, the only criterion that uses the EPV measure, which is contrasted with the fact that many other criteria use the FI (Veerkamp; Berger, 1997; Han, 2018).

The Likelihood Weighted Information (LWI) is a source of information derived from the FI with an additional source of data: the likelihood function (Veerkamp; Berger, 1997). In a study devised by its creators, the Maximum LWI criterion (MLWI) was shown to be better than the MI at choosing an item that will optimize ability estimates. However, further research has struggled to reproduce these findings (Penfield, 2007; Reeve, 2007; Choi; Schwarz, 2009). The LWI at item *j* for ability θ is the Fischer information multiplied by the likelihood of θ given responses $x_1, \dots, x_k$, where k is the item that has just been answered (Veerkamp, & Berger, 1997). It is displayed in Equation 4.

$$LWI_j(\theta) = \int_{-\infty}^{\infty} I_j(\theta) \cdot L(\theta \mid x_1, \dots, x_k) d\theta \tag{4}$$

That same logic can be employed in a Bayesian approach that was also presented by Veerkamp and Berger (1997). The Posterior Weighted Information (PWI) weighs the information function by the posterior, meaning that, in addition to the likelihood function, it also takes the prior into account. So, the PWI at item *j* for ability θ is also the Fisher information multiplied by the likelihood function, but it includes a prior function π at ability θ in the calculation (van der Linden, 1998). Its formula is displayed in Equation 4.

$$PWI_j(\theta) = \int_{-\infty}^{\infty} I_j(\theta) \cdot \pi(\theta) \cdot L(\theta \mid x_1, \dots, x_k) d\theta \tag{5}$$

The criterion that employs the PWI, the Maximum PWI (MPWI), has also been shown to be superior to MI at choosing an item that will optimize ability estimates (van der Linden; Pashley, 2000). But, once again, other researchers struggled to reproduce these findings (Penfield, 2007; Reeve, 2006; Choi; Schwarz, 2009).

An additional, newer, measure, the Expected Information (EI), reproduces the technique of weighing the Fisher information, but its authors used, as weight, the theta estimation function (Han, 2018). The EI measure for ability θ at item *j* is the Fisher information multiplied by the probability of ability θ given a prior with mean μ and standard deviation σ^2 (Han, 2018).

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5, n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

This can be seen in Equation 6.

$$EI_j(\theta) = \int_{-\infty}^{\infty} I_j(\theta) \cdot p(\theta \mid \mu, \sigma^2) d\theta \quad (6)$$

The performance of these criteria have been compared in multiple studies. Choi and Swartz (2009) ran a study which compared the effectiveness of MFI, MLWI, MPWI, MEPV, MEI, and random selection methods. The authors also examined the Maximum Expected Posterior Weighted Information (MEPWI), which they found to be mathematically identical to the MPWI. The results of van der Linden and Pashley (2000) were put into question, as they had suggested that the MEPWI had been statistically superior to the MPWI.

In any case, the authors found that the performance of all methods, except for the random, were similar, which also comes into conflict with results from Veerkamp and Berger (1997) which had found the MI to be inferior to the MLWI and MPWI (Choi; Swartz, 2009). The reason for this conflict may be related to the characteristics of the tests. While Verkamp and Berger (1997) used educational test data, Choi and Swartz (2009) used a quality of life scale. No studies have compared these different criteria in psychological tests.

1.2 PSYCHOLOGICAL TEST CATS

Few psychological tests based on CAT report the criterion they use to select items adaptively. When they do report, comparisons are not typically made. Chang (2009) reported the development of a cognitive diagnosis CAT along with two metrics to be used as criteria for item selection. The author compared these criteria with algorithms based on Kullback-Leibler divergence (Cover; Thomas, 2012; Kullback; Leibler, 1951), which, applied to CAT, became known as the Kullback-Leibler information (KL), and with algorithms based on Shannon entropy (Shannon, 1948). They found that the two new criteria had improved performance in some, but not all, situations. However, these algorithms were compared to a negative control, the random algorithm. The author did not compare these new criteria with any of the criteria which uses the FI, such as the MFI, MLWI, MPWI and MEI.

In Brazil, a systematic review by Peres (2019) revealed that few CAT experiences have been reported. Most experiences were dissertations that employed CAT in educational assessment. Only two studies report the use of CAT in psychology: one dedicated to the screening of dyslexia (Santos, 2017), and other to create an item bank for assessing the Big Five factors of personality (Oliveira, 2017). However, the study by Santos (2017) did not actually employ any CAT methodology. Meanwhile, Oliveira (2017) used the Concerto platform to test 525 items through an incomplete block design. The final item bank was composed of 317 items.

Since this study aimed to create an item bank for a CAT, there were no indications that an ISA was used, since using incomplete block designs with the specific goal of calibrating all items would require specific items to be administered, making it incompatible with the use of an ISA. The criteria used by ISA have had different performance in psychological and educational tests, but this has not been studied further. In fact, no psychological tests were found to use the CAT format at all in Brazil. The present study sought to examine whether two EI tests originally developed by MacCann and Roberts (2018) and then adapted to a Brazilian Portuguese audience by Souza (2023) could be efficiently administered as a CAT without impact to score validity. In this study, we aimed to assess the different criteria employed by the ISA to elucidate whether a computerized, adaptive administration of these tests has advantages compared to the traditional form. We also aimed to compare the performance of the criteria to find out the ideal settings for a CAT.

2 METHODS

2.1 PARTICIPANTS

The study utilized the same dataset from the EI tests' validation study. The sample comprised 688 participants, overwhelmingly female (81.5%), undergraduate students (21.7%) and single (54.8%). The median age was 23 (MAD = 7.4; Revelle, 2023). Participants were recruited by means of a social media campaign.

2.2 INSTRUMENTS

The Concerto Platform (University of Cambridge Psychometrics Center) and Google Forms (Alphabet Inc.) were utilized for data collection. A Free Consent and a demographic form were also utilized. In total, four psychological tests were utilized. These included the Brazilian Portuguese adaptations of both the Situational Test of Emotional Understanding and Emotional Management (Souza, 2023), as well as two scales that were used in a different study (Souza, 2023): the Reduced Scale of the Big Five Personality Factors (Passos; Laros, 2015) and the Life Satisfaction Scale (Oliveira et al., 2009).

The Situational Test of Emotional Understanding (STEU; Souza, 2023; MacCann; Roberts, 2008) is a 32-item multiple-choice assessment that measures an individual's ability to identify emotions in context. Each item presents a description of an emotionally charged situation involving a fictitious character, and the respondent must identify the emotion that the character is most likely to experience in that scenario. For each item, only one alternative is correct. Outcomes are correct or wrong. Item response theory (2-parameter model; Lord; Novick, 1968) fit indices were favorable under cutoff values proposed by Cai and Hansen (2013) and Hu and Bentler (1999), M^{2*} (432) = 749.62, $p < .001$, RMSEA = .033 [CI 95 .029; .037], SRMSR = .045, NNFI = .0948, CFI = .955.

The Situational Test of Emotional Management (STEM; Souza, 2023; MacCann; Roberts, 2008) uses 30 multiple-choice items to measure whether individuals can identify among the presented options, the most effective response to an emotionally charged situation. For the STEM, answer outcomes may be correct or wrong, but multiple answers can also have different scores.

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5, n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

Fit indices for item response theory (generalized partial credit model; Muraki, 1992) were also favorable, M^{2*} (375) = 408.517, p < .001, RMSEA = .011 [CI95 0; .018], SRMSR = .045, NNFI = .998, CFI = .998.

Pearson correlation between the scores calculated from the STEU and the STEM are .501.

The R programming language (version 4.2.2) was utilized for data analysis. CAT simulations were performed using the mirtCAT package (Chalmers, 2016), and visualizations were created with the ggplot2 package (Wickham, 2016). All code used in this study was original and can be provided upon request.

2.3 PROCEDURE

The study was approved by the Research Ethics Committee of the University of São Paulo (approval number 1.780.012).

Recruitment for the study was conducted through social media platforms, where participants were provided with a link to the testing platforms. Data collection initially took place on the Concerto Platform but was later switched to Google Forms due to hosting issues. The forms and tests were presented on separate pages and only on the Google Forms platform could participants resume previous sessions. Participants could only submit the form when it had been fully filled out.

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5 n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

2.4 ANALYSIS

2.4.1 CAT SIMULATIONS

Two different simulation studies were run. Their details are displayed in Box 1.

Box 1 – Input parameters of Computer-Adaptive Testing simulation studies 1 and 2.

Characteristics	Study 1	Study 2
Item parameters	The item parameters for both the STEU and the STEM were calibrated in Souza (2023).	
Algorithms	MFI, MLWI, MPWI, MEI, IKL, IKLn, IKLP, and IKLPn	
Stopping Rule	$\Delta SE < .001$	Fixed number of items
Repetition	One simulation per algorithm per test.	k simulations per algorithm per test, where k = number of items per test.
Number of simulations	1 simulation per 11 algorithms per each of the two tests, $11 \times 2 = 22$ simulations	32 simulations for the STEU and 30 simulations for the STEM per 11 algorithms, $32 \times 11 + 30 \times 11 = 682$ simulations
Analysis	Pearson correlation between theta estimate per algorithm and the theta estimate with the entire test. Frequency of number of items required by each algorithm to reach the stopping rule for each participant.	Mean standard error levels at each number of items administered.
Purpose	Bias Test length reduction	Precision estimate Test length reduction

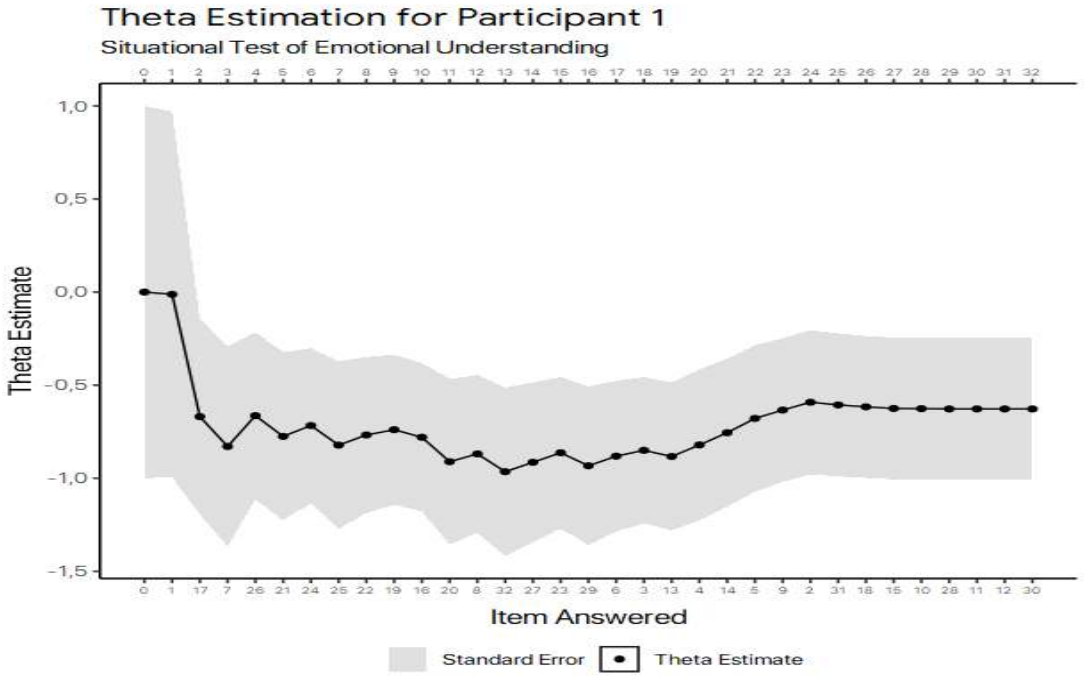
Source: Built by the authors.

Although the exact format of analysis is novel, the statistical tests did not differ in principle from previous studies that compared the efficacy and efficiency of item selection criteria. They involve comparing the final theta estimate of the simulation to the known results of the full score, as a measure of algorithm efficacy, and comparing the number of items required to reach that estimate among different criteria, as a measure of algorithm efficiency (e.g., Chang, 2009; Choi; Schwarz, 2009). The SE stopping rule criterion is defined as the industry standard, which have widespread use on statistical software (e.g., Chalmers, 2016).

For both studies, the mirtCAT package was used to generate answers based on the empirical data response pattern collected during the validation study, simulating new data collection using an ISA for each participant. The IRT parameters were also calibrated in the validation study. In the first study, the simulation procedure was repeated once for each ISA for each of the two tests.

The algorithms that were tested were the maximum Fisher information (MFI), the minimum expected posterior variance (MEPV), the maximum likelihood-weighted information (MLWI), the maximum posterior-weighted information (MPWI), the maximum expected information (MEI), and the integration-based Kullback–Leibler criteria with or without prior density, and with or without root-n Weight (corresponding to algorithms IKLPn, IKLP, IKLn and IKLn, where the "P" denotes the prior density weight and, the "n", the root-n weight).
Each simulation study yielded one or more databases containing data for each of the participants as if they had answered a CAT test, a database containing the items that they answered until they reached the stopping rule, the estimated theta values after each item answered, the standard error (SE) of each estimate, and the final theta estimate. An example of simulated test administration can be found in Figure 1.

Figure 1 – Theta estimates and standard errors after each response from one simulated participant.



For each ISA, the correlation between the final theta found and the estimated theta was calculated. The same was done between the estimated SE and the number of items administered.

For the second study, each algorithm was simulated *k* times, where *k* was the test's maximum number of items (32 for the STEU and 30 times for the STEM). This was necessary because this study employed a different stopping rule: the test would stop when it administered a fixed number of items, at every number between one and the number of items of the test. This made it possible to calculate the mean SEs for each ISA at each fixed number of items administered, simulating alternative testing conditions, such as content constraints. The standard deviations of the SEs were also calculated.

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5 n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

3 RESULTS

All ISAs reached a correlation greater than $r = .98$ between the theta estimates they produced and the theta estimates calculated using the entire test. The correlation between the number of items required for the test to end and the SE was also calculated, and was significant for all ISAs, for both tests, ranging between .412 and .71, which are considered medium-sized and high (Cohen, 1992). These statistics are shown in Table 1.

Table 1 – Correlation between abilities estimated utilizing item selection algorithms and abilities estimated using all test items.

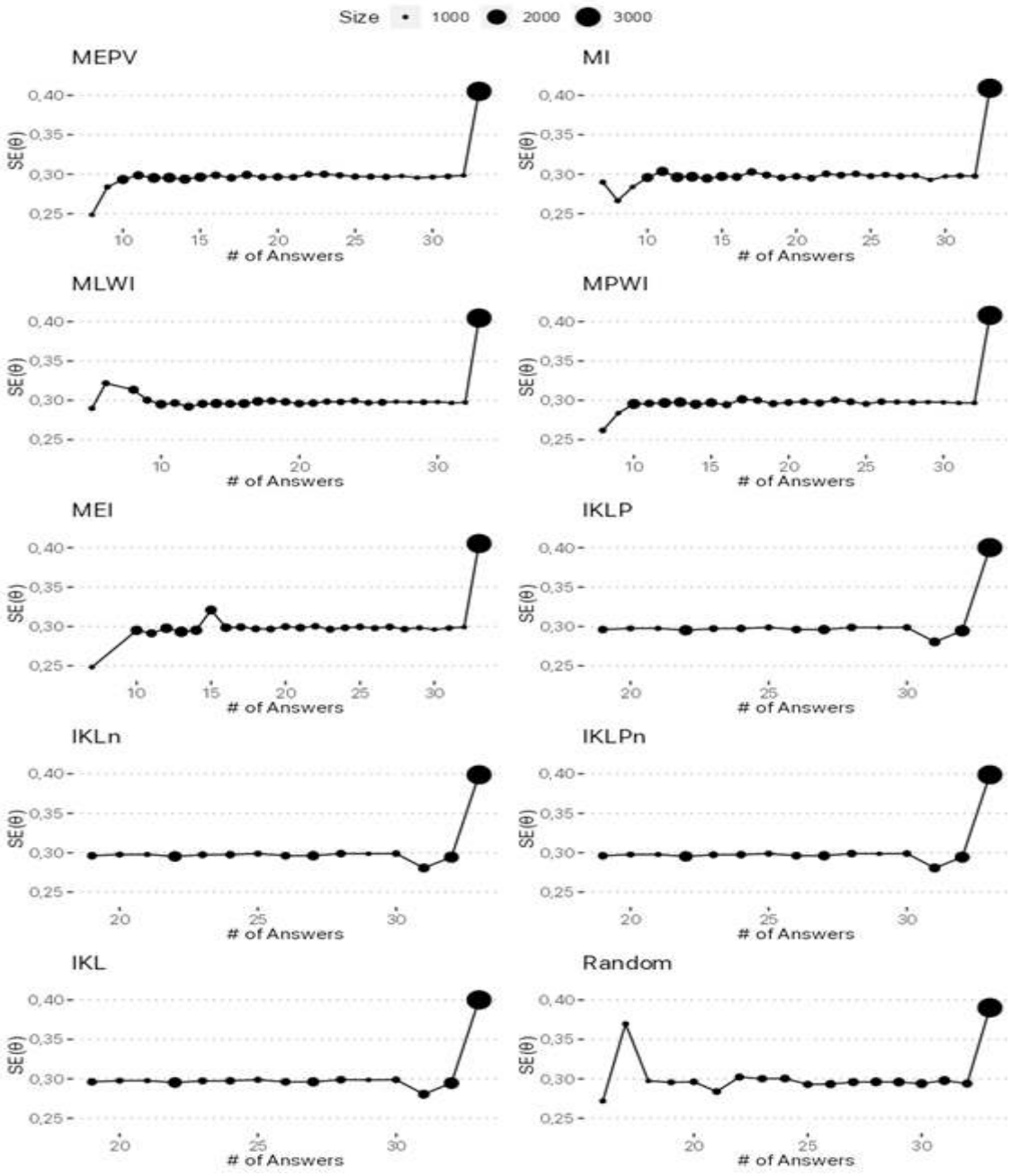
	STEU		STEM	
	Correlation		Correlation	
Algorithm	Theta	N-SE	Theta	N-SE
MFI	.982	.601	.996	.624
MEPV	.989	.594	.994	.614
MLWI	.985	.580	.994	.606
MPWI	.988	.595	.993	.611
MEI	.983	.599	.994	.607
IKL, IKLn, IKLp, IKLPn	.994	.423	.998	.710
Random	.992	.435	.996	.573
Sequential	.994	.421	.998	.710

Notes. MFI = maximum Fisher information, MEPV = minimum expected posterior variance, MLWI = maximum likelihood-weighed information, MPWI = maximum posterior-weighted information, MEI = maximum expected information, IKL = integration-based Kullback–Leibler criteria, IKLn = IKLn with root-n weight, IKLP = IKLP with prior density weight, IKLPn = IKLP with root-n weight and prior density weight. All KL-based algorithms yielded the same correlation sizes.

As shown in Figure 2, the overall SE outcomes of the ISAs were close between the FI-based algorithms MEPV, MFI, MLWI and MPWI, and the KL-based algorithms. This is expected, as the stopping rule is designed for the SE to be minimized—however, the KL-based algorithms administered more items to reach these low SE levels. In general, the pattern in the plot indicates that when item information matches the participants' ability, the test ends quickly with low SE. Until the ISA had to serve all items, information about the participants' proficiency was able to reach the required $\Delta SE < 10^{-1}$ criterion. When all items are served, that means the criterion was not met, but this does not limit the SE; it can vary freely. Indeed, different ISAs have different maximum SEs.

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5, n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

Figure 2 – Association of the mean standard error and number of responses of the item selection algorithms.



Frequency statistics for the number of items required to reach the stopping rule are shown in Tables 3 and 4 for the STEU and STEM tests, respectively. For the STEU, the KL-based algorithms did not reach the required $\Delta SE < .001$ rule (Chalmers, 2016) to end the test for any simulated participant before item 19.

This contrasts with the MFI, MEPV, MLWI, MPWI and MEI methods, which reached that requirement for between 37.64% and 40.12% simulated participants before item 19, including

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5 n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

between 5.32% and 10.17% estimations of simulated participants reaching the requisite SE difference rule before item 11. All algorithms had simulated participants reach the stopping rule at least once before the last item of the test, with the number of participants varying per algorithm—the number ranged from between 299 for the random algorithm to 368 for the MLWI algorithm. Among the adaptive algorithms, the worst performance was a tie between the IKLn and IKLPn ISAs, with 326 simulated participants reaching the stopping rule before reaching the last item of the test.

Table 2 – Number of items administered by each algorithm for the situational test of emotional understanding.

	Number of Items Administered					
Algorithm	4-9	10-17	18-24	25-31	1-31	32
IKL	0	0	124	207	331	357
IKLn	0	0	124	202	326	362
IKLP	0	0	124	207	331	357
IKLPn	0	0	124	202	326	362
MEI	37	230	67	26	360	328
MEPV	36	237	60	19	352	336
MFI	40	236	63	29	368	320
MLWI	70	189	73	24	356	332
MPWI	54	218	66	24	362	326
Random	0	5	82	212	299	389
Total	237	1115	907	1152	3411	3469

Notes. Sample sizes = 688. Item number grouping chosen to highlight differences.

The STEM simulations were less successful in reducing the number of items needed to reach the stopping rule, but test lengths were still significantly reduced. The four KL-based algorithms again had identical performance and only started reaching the stopping rule when 23 items were administered. This contrasts with the MFI, MEPV, MLWI, MPWI and MEI methods, which reached the stopping rule for around 22% of simulated testing sessions before item 23.
Once again, all algorithms had simulated participants reach the stopping rule at least once before the last item of the test. The number of such participants ranged from 172 for the random algorithm to 187 for the MLWI algorithm.

Notably, for the STEM, the worst performance was a tie between the KL-based ISAs, which performed worse than the random algorithm.

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5, n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

Table 3 – Number of items administered by each algorithm for the situational test of emotional management.

	Number of Items Administered			
Algorithm	3–22	23–29	1–29	30
IKL	0	169	169	519
IKLP	0	169	169	519
IKLPn	0	169	169	519
IKLn	0	169	169	519
MEI	153	33	186	502
MEPV	155	28	183	505
MFI	156	31	187	501
MLWI	154	31	185	503
MPWI	155	32	187	501
Random	113	59	172	516

Notes. Sample sizes = 688. Item number grouping chosen to highlight differences.

The efficient performance of FI-based algorithms was also observed on Study 2. For the STEU, the best ISAs reached the minimum mean SE on simulations with 26 items. This indicates that, on average, precision is already at its highest value for the best ISAs even before the last six items had been administered. Comparatively, the same mean SE level is only reached by the KL-based algorithms when simulations had administered all items. The MEPV did not perform correctly, as mean SE estimates varied back and forth. These results can be seen in Figure 3.

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5 n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

Figure 3 – Mean standard error per number of items for each item selection algorithm for the situational test of emotional understanding, represented with colors set linearly.



Standard Error - Mean

	MI	MEPV	MLWI	MPWI	MEI	IKL	IKLP	IKLn	IKLPn	Random
1	1,016	1,016	1,016	1,016	1,016	1,016	1,016	1,016	1,016	1,016
2	0,680	0,798	0,748	0,680	0,773	0,778	0,778	0,778	0,778	0,904
3	0,635	0,597	0,658	0,610	0,550	0,779	0,779	0,779	0,779	0,841
4	0,386	0,496	0,606	0,467	0,491	0,638	0,638	0,638	0,638	0,770
5	0,441	0,470	0,516	0,455	0,453	0,584	0,584	0,584	0,584	0,688
6	0,423	0,448	0,464	0,436	0,423	0,582	0,582	0,582	0,582	0,653
7	0,406	0,433	0,438	0,414	0,409	0,536	0,536	0,536	0,536	0,615
8	0,394	0,424	0,419	0,392	0,395	0,530	0,530	0,530	0,530	0,575
9	0,384	0,398	0,405	0,386	0,388	0,516	0,516	0,516	0,516	0,570
10	0,377	0,411	0,393	0,379	0,381	0,505	0,505	0,505	0,505	0,539
11	0,372	0,377	0,383	0,375	0,374	0,504	0,504	0,504	0,504	0,535
12	0,367	0,400	0,378	0,368	0,370	0,496	0,496	0,496	0,496	0,498
13	0,362	0,366	0,370	0,364	0,366	0,493	0,493	0,493	0,493	0,512
14	0,358	0,400	0,365	0,361	0,365	0,478	0,478	0,478	0,478	0,479
15	0,356	0,358	0,363	0,360	0,361	0,469	0,469	0,469	0,469	0,487
16	0,352	0,390	0,357	0,356	0,358	0,446	0,446	0,446	0,446	0,450
17	0,349	0,353	0,355	0,355	0,355	0,407	0,407	0,407	0,407	0,467
18	0,347	0,384	0,353	0,352	0,352	0,392	0,392	0,392	0,392	0,433
19	0,345	0,350	0,352	0,349	0,350	0,392	0,392	0,392	0,392	0,447
20	0,345	0,383	0,352	0,348	0,348	0,390	0,390	0,390	0,390	0,413
21	0,346	0,349	0,350	0,347	0,346	0,378	0,378	0,378	0,378	0,439
22	0,346	0,382	0,348	0,348	0,346	0,375	0,375	0,375	0,375	0,402
23	0,346	0,347	0,348	0,347	0,346	0,372	0,372	0,372	0,372	0,420
24	0,348	0,377	0,350	0,349	0,346	0,369	0,369	0,369	0,369	0,388
25	0,345	0,345	0,346	0,346	0,345	0,367	0,367	0,367	0,367	0,410
26	0,344	0,377	0,346	0,345	0,344	0,360	0,360	0,360	0,360	0,370
27	0,343	0,344	0,344	0,344	0,344	0,358	0,358	0,358	0,358	0,404
28	0,343	0,378	0,344	0,344	0,343	0,358	0,358	0,358	0,358	0,361
29	0,343	0,344	0,344	0,344	0,344	0,356	0,356	0,356	0,356	0,393
30	0,343	0,377	0,344	0,344	0,344	0,353	0,353	0,352	0,352	0,354
31	0,343	0,344	0,344	0,344	0,344	0,346	0,346	0,346	0,346	0,384
32	0,344	0,378	0,344	0,344	0,344	0,344	0,344	0,344	0,344	0,344

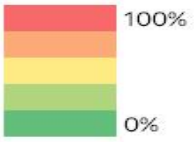
For the STEM, mean SEs took longer to stabilize. The MEI, MFI, MLWI, and MPWI methods reached the stopping rule only when 29 items had been administered. However, the MEPV algorithm did not exhibit the erratic behavior observed on the STEU. Meanwhile, the IKL methods did not stabilize at a low SE pattern until simulations were done with all test items. These results can be seen in Figure 4.

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5, n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

Figure 4 – Mean standard error per number of items for each item selection algorithm for the situational test of emotional management, represented with colors set linearly.

Standard Error - Mean

	MI	MEPV	MLWI	MPWI	MEI	IKL	IKLP	IKLn	IKLPn	Random
1	0,887	0,887	0,887	0,887	0,887	0,887	0,887	0,887	0,887	0,887
2	0,752	0,742	0,762	0,777	0,742	0,828	0,828	0,828	0,828	0,815
3	0,637	0,651	0,720	0,665	0,669	0,759	0,759	0,759	0,759	0,758
4	0,585	0,593	0,631	0,603	0,620	0,713	0,713	0,713	0,713	0,710
5	0,547	0,555	0,576	0,564	0,569	0,689	0,689	0,689	0,689	0,670
6	0,519	0,525	0,544	0,533	0,534	0,654	0,654	0,654	0,654	0,634
7	0,496	0,503	0,518	0,508	0,507	0,625	0,625	0,625	0,625	0,607
8	0,477	0,483	0,495	0,489	0,486	0,591	0,591	0,591	0,591	0,582
9	0,462	0,467	0,477	0,471	0,471	0,568	0,568	0,568	0,568	0,560
10	0,449	0,453	0,462	0,456	0,457	0,556	0,556	0,556	0,556	0,539
11	0,439	0,441	0,448	0,445	0,444	0,523	0,523	0,523	0,523	0,524
12	0,429	0,431	0,437	0,434	0,433	0,510	0,510	0,510	0,510	0,506
13	0,419	0,422	0,427	0,424	0,424	0,493	0,493	0,493	0,493	0,490
14	0,411	0,413	0,418	0,415	0,415	0,478	0,478	0,478	0,478	0,480
15	0,403	0,406	0,409	0,407	0,407	0,461	0,461	0,461	0,461	0,466
16	0,396	0,399	0,402	0,400	0,400	0,448	0,448	0,448	0,448	0,454
17	0,390	0,392	0,395	0,393	0,393	0,439	0,439	0,439	0,439	0,443
18	0,385	0,387	0,389	0,387	0,388	0,431	0,431	0,431	0,431	0,436
19	0,379	0,381	0,383	0,382	0,383	0,415	0,415	0,415	0,415	0,426
20	0,374	0,376	0,378	0,377	0,377	0,412	0,412	0,412	0,412	0,415
21	0,370	0,372	0,374	0,372	0,373	0,408	0,408	0,408	0,408	0,408
22	0,367	0,368	0,369	0,369	0,368	0,400	0,400	0,400	0,400	0,400
23	0,363	0,364	0,366	0,365	0,365	0,385	0,385	0,385	0,385	0,394
24	0,361	0,361	0,362	0,362	0,361	0,377	0,377	0,377	0,377	0,386
25	0,358	0,359	0,359	0,359	0,359	0,371	0,371	0,371	0,371	0,381
26	0,356	0,357	0,357	0,357	0,357	0,370	0,370	0,370	0,370	0,374
27	0,354	0,355	0,355	0,355	0,355	0,363	0,363	0,363	0,363	0,368
28	0,353	0,353	0,353	0,353	0,353	0,359	0,359	0,359	0,359	0,362
29	0,352	0,352	0,352	0,352	0,352	0,355	0,355	0,355	0,355	0,357
30	0,351	0,351	0,351	0,351	0,351	0,351	0,351	0,351	0,351	0,351



Additionally, mean SEs for experimental ISAs are significantly lower than the negative control (the random algorithm), even from simulations with just two items. In other words, any algorithm that made use of information—whether FI or KL and whether weighted or not—was more effective than the algorithm that randomly selected items, even if only one item is selected.

4 DISCUSSION

This study aimed to determine whether the Brazilian versions of the STEU and the STEM could be efficiently administered using a CAT mechanism without loss of validity and precision. Given the relationship between increased measurement validity in CAT and reduced testing times. To this end, we employed a methodology based on simulated administrations of real-world data for which test results were previously known. This allowed us to assess the extent to which administering fewer items could accurately estimate proficiency levels while also including the measurement error expected in the response patterns collected in a typical test administration session.

In the first study simulations were run without establishing a fixed number of items that would be administered to each participant.

Different ISAs could administer whichever items they chose, and the number of items administered would be a consequence of the stopping rule being reached, which was the same for all simulations ($\Delta SE < 10^{-3}$). Using this method, it was possible to determine that the

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5 n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

correlation between the thetas estimates in the simulations and those estimated using the entire test was over .9 for all ISAs.

Assuming that reducing testing times would be one of the main goals of this step, the homogeneity among the correlation between thetas could hide different performances. For example, even the random algorithm reached elevated levels of correlation, outperforming several other algorithms. This measure alone is inadequate to measure the ISAs' performance, since the values of these correlations does not take into account whether an algorithm's ability to reach ability estimates so close to those obtained using full information was due to its simulations requiring the administration of additional items. As such, this did not yet allow for the study of the expected SE for each number of items administered for every participant response pattern.

To address this issue, the second study involved constraining analyses to an exact number of items administered for each ISA. Analyzing the results of this study helped determine that the seemingly positive results of the random algorithm are associated with it requiring the administration of many more items to reach the stopping rule. In this step, every single ISA had significantly better mean SEs, from the first item they selected all the way to the one before the last. At this point, every ISA, including the random algorithm, reached the minimum SEs possible for the test they were simulating, since all items had been administered. Indeed, the random algorithm could only reach the minimum SEs for all simulated participants at that point.

Considering that the ultimate goal of an examination is to estimate the examinee's latent trait, the results of the this study suggest that it is possible to achieve optimal, or optimal or near-optimal mean SE levels with significantly fewer items administered in either of the tests, provided that informative ISAs are used. Using the $\Delta SE < 10^{-2}$ stopping rule, which is an often acceptable setting, as many as 15 and 6 fewer items could be administered by the best-performing ISAs in the STEU and the STEM, respectively. This represents a reduction of 21 items out of 62, or 34% of total items answered.

This analysis helps to understand the impact of the mean SE. However, in practical terms, without any loss to the SE, the first study already revealed that more than half of all participants could be expected to need to answer at least one fewer item for the most efficient ISAs. For the MFI algorithm, this was 368 participants, versus 320 who had to answer all items.

In comparing these results with other research, our data does not seem to agree with papers by the authors of nearly every algorithm competing with the MFI algorithm (e.g., van der Linden, 1998; van der Linden; Pashley, 2000; Veerkamp; Berger, 1997). Instead, our findings are consistent with those reported by Choi and Swartz (2009), who found that the MFI, MLWI, and MPWI were roughly close in performance. Excluding the its SE performance at every second item, the MEPV algorithm performed as efficiently as the best performing algorithm, the MFI.

A study that compared the MFI, MPWI and MEPV using testlets also reached the conclusion that these algorithms' performance is overall similar (Murphy; Dodd; Vaughn, 2010)

Indeed, considering Choi and Schwartz's (2009) finding that the MEPWI was mathematically identical to the MPWI despite previous findings that it was worse than the MPWI, it is possible that differences in the transformation of the mathematical formulae into computer algorithms could account for any reported differences in the literature. We hypothesize that a similar issue may be responsible for the issue with the ability estimate at every other item found in the MEPV results for the present study.

Methods such as the MEI (Han, 2018), MLWI, and MPWI (van der Linden, 1998) all work similarly, making use of the FI weighed by a probability or likelihood function and possibly a prior. The MLWI algorithm weighs the same information by the likelihood of theta given the response pattern and can be viewed in Equation 3. Given that all the prior for all calculations was a normal distribution N (0;1), it is expected that these methods would have similar performance. Further studies should examine the performance of these methods under different priors.

Finally, the algorithms based on the KL had the poorest mean SE performance among the non-random ISAs. The IKL algorithm (Cover; Thomas, 1991) was developed to address situations for which the FI would not be adequate (Chang; Ying, 1996). Specifically, its use is recommended during testing at points in which the test's current ability estimate is not necessarily likely to be the true theta, such as at the start of the test, when a generic prior is utilized. This has been implemented by estimating global information at each item for KL-based algorithms, whereas the MFI algorithm is based directly on how much local information each item carries (Chang; Ying, 1996)

Local information refers to the amount of information that items provide at theta values close to the current theta estimate. In contrast, global information encompasses all the information that items provide, including information farther from that point. These concepts are similar and there is a mathematical relationship between them, as shown in Equation 7 (Dabak; Johnson, 2003).

$$I_{\mathrm{j}}(\theta) = \frac{\partial^2}{\partial \theta^2} \mathrm{K}^{(\mathrm{n})}(\theta||\theta)|_{\theta-\theta_0} \qquad (7)$$

Where $K^{(n)}(\theta||\theta)$ is the KL.

The MFI is defined as the inverse of the expected value of the second derivative of the log-likelihood function with respect to theta (Ly et al., 2017). In contrast, when applied to the item selection (Cover; Thomas, 1991), the Kullback and Leibler (1951) divergence is the expected value of the logarithmic difference between the likelihood and the theta estimation function (Dabak; Johnson, 2003).

Therefore, the KL is the second derivative of the FI. The two will be equal when the KL is calculated for a small interval of theta. Since KL considers information for a larger interval, the MFI's performance will be better when the test taker's true ability is close to the ability estimate. In this way, the KL's focus on global information may lead it to select items that contain information not relevant to the test's theta estimate of the test taker.

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5 n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

However, Chang and Ying's (1996) study revealed that KL-based algorithms performed better in terms of mean squared error and bias under many circumstances. In their study, mean SE was not measured; instead, mean squared error and bias were calculated in relation to the true theta score. In our study, correlations between the various ISA and the true theta score revealed that KL-based algorithms had marginally better bias (mean $r = .994$) when compared to the FI-based algorithm that had the highest correlation ($r = .989$).

Another factor that may have influenced the results is the stopping rule used in this study. We adopted the default stopping rule of $\Delta SE < 10^{-3}$, which was also used in Choi and Swartz's (2009) study. The use of a ΔSE criterion may have contributed to the superior performance of the MFI algorithm. While some studies have examined the effect of stopping rules (e.g., Babcock; Weiss, 2012), no study has assessed their potential effect on ISAs. Future studies that examine ISA performance under different stopping rules, such as those based on the difference between ability estimates, are therefore warranted.

Additionally, this study did not attempt to use content balancing rules. However, their effect is largely independent from the performance of the ISAs except insofar as they control the minimum number of items that must be administered.

There are several limitations to this study. For instance, although using empirical data as response patterns for the simulations has been considered a novel means of accruing justification for further CAT implementations of the STEU and the STEM, no study has investigated whether there is a negative impact of doing so with the same data utilized in parameter calibration. While the utilization of cross-validation methods has been considered largely superseded by IRT fit measures such as the M^2* (Cai; Hansen, 2013) and C^2 (Cai; Monro, 2014) model fit measures, the S-χ^2 (Kang; Chan, 2007) item fit measure, and the traditional measures Iz and Zh being utilized for person fit (Drasgow et al., 1985; Felt et al., 2017), all of which were employed in the validation study for the STEU and the STEM (Souza, 2023), previous studies have suggested various situations in which repeating samples would be considered a limitation (de Rooiji; Weeda, 2020). For instance, Cunha (2019) showed that constructing a prediction model based on the same data in which the predictor was produced led to estimates below the expected risk value. While the present study did not employ any further parameter calculation using the same covariance matrix, it is possible that some other problem may have arisen.

With respect to making sure that the tests benefit from CAT's many reported benefits (Souza, 2023), one important limitation is the size of the item bank. Several benefits have been associated with large item pools, such as the reduction of the exposure effect, which reduces the bias associated with individuals taking the same test more than once.

However, since both the items and IRT parameters have been published (Souza, 2023), further studies may construct additional items which can be easily calibrated using incomplete block designs (Ariel; van der Linden; Veldkamp, 2006).

In summary, regardless of individual ISA performance, to the extent to which the goal of administering the Brazilian adaptation of the STEU and the STEM is the estimation of the latent traits emotional understanding and emotional management, these results show that the

RECITAL - Revista de Educação, Ciência e Tecnologia de Almenara, v. 5, n. 3, set/dez 2023.
ISSN: 2674-9270. Disponível em: http://recital.almenara.ifnmg.edu.br/index.php/recital

CAT versions of the STEU and the STEM were able to estimate these abilities with the same precision and validity as the completed test versions while benefitting from the advantages CAT confers. This development should allow researchers to achieve increased measurement validity while reaching more participants by offering more attractive research participation opportunities with lower testing times. Finally, this study employed a novel methodology for seeking evidence of benefits that test developers stand to gain when adapting tests to CAT. We hope that psychological test developers will take notice and employ similar techniques.

REFERENCES

ARIEL, A; VAN DER LINDEN, W. J; VELDKAMP, B. P. A Strategy for Optimizing Item-Pool Management. **Journal of Educational Measurement, v.** 43, n. 2, p. 85–96, 2006. https://doi.org/10.1111/j.1745-3984.2006.00006.x

BABCOCK, B; WEISS, D. J. Termination Criteria in Computerized Adaptive Tests: Do Variable-Length CATs Provide Efficient and Effective Measurement? **Journal of Computerized Adaptive Testing,** v. 1**,** n. 1–5, p. 1–18, 2012. https://doi.org/10.7333/1212-0101001

BOCK, R. D; MISLEVY, R. J. Adaptive EAP estimation of ability in a microcomputer environment. **Applied Psychological Measurement,** v. 6**,** n. 4, p. 431–444, 1982. https://doi.org/10.1177/014662168200600405

CAI, L; HANSEN, M. Limited-information goodness-of-fit testing of hierarchical item factor models. **British Journal of Mathematical and Statistical Psychology**, v. 66, p. 245–276, 2013. https://doi.org/10.1111/j.2044-8317.2012.02050.x

CHOI, S. W; SWARTZ, R. J. Comparison of CAT Item Selection Criteria for Polytomous Items. **Applied Psychological Measurement,** v. 33**,** n. 6, p. 419–440, 2009. https://doi.org/10.1177/0146621608327801

CHANG, H.-H; YING, Z. A global information approach to computer-adaptive testing. **Applied Psychological Measurement,** v. 20**,** n. 3, p. 213–229, 1996. https://doi.org/10.1177/014662169602000303

CHENG, Y. When cognitive diagnosis meets computer-adaptive testing: CD-CAT. **Psychometrika,** v. **74,** n. 4, p. 619–632, 2009. https://doi.org/10.1007/s11336-009-9123-2

COHEN, J. A power primer. **Psychological bulletin**, v. 112, n. 1, p. 155, 1992.

COVER, T. M; THOMAS, J. A. **Elements of information theory**. 2 ed. Wiley, 2012.

DABAK, A. G; JOHNSON, D. H. **Relations between Kullback-Leibler distance and Fisher information**. Rice University, 2003. http://dhj.rice.edu/files/2014/07/distance.pdf

DE ROOIJ, M; WEEDA, W. Cross-validation: A method every psychologist should know. **Advances in Methods and Practices in Psychological Science,** v. 3**,** n. 2, p. 248–263, 2020. https://doi.org/10.1177/2515245919898466

FELT, J. M; CASTANEDA, R; TIEMENSMA, J; DEPAOLI, S. Using Person Fit Statistics to Detect Outliers in Survey Research. **Frontiers in Psychology,** v. 8, p. 863, 2017. https://doi.org/10.3389/fpsyg.2017.00863

FIORI, M.; AGNOLI, S.; DAVIS, S. K. New trends in emotional intelligence: conceptualization, understanding and assessment. **Frontiers in Psychology**, v. 14, Article 1266076, 2023. https://doi.org/10.3389/fpsyg.2023.1266076

FRIEDEN, B. R; GATENBY, R. A. Principle of maximum Fisher information from Hardy's axioms applied to statistical systems. **Physical Review E,** v. 88**,** n. 4, Article 042144. https://doi.org/10.1103/PhysRevE.88.042144

HAN, K. Components of the item selection algorithm in computerized adaptive testing. **Journal Of Educational Evaluation for Health Professions**, v. 15**,** n. 1, Article 7, 2018. https://doi.org/10.3352/jeehp.2018.15.7

HU, L.; BENTLER, P. M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives". **Structural Equation Modeling**, v. 6, n. 1, p. 1–55, 1999. https://doi.org/10.1080/10705519909540118

KULLBACK, S; LEIBLER, R. A. On information and sufficiency**. Annals of Mathematical Statistics,** v. **22,** n. 1, p. 79–86, 1951. https://doi.org/10.1214/aoms/1177729694

LING, G; ATTALI, Y; FINN, B; STONE, E. A. Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? **Applied Psychological Measurement,** v. 41**,** n. 7, p. 495–511, 2017. https://doi.org/10.1177/0146621617707556

LORD, F. M; NOVICK, M. R. **Statistical theory of mental test scores**. Addison-Wesley, 1968.

LORD, F. M. **Applications of Item Response Theory to Practical Testing Problems**. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203056615

LUECHT, R; SIRECI, S. **A Review of Models for Computer-Based Testing** (ERIC Document № ED465763), ERIC Database, 2011.

LUECHT, R. M. Computer-Adaptive Testing. **Wiley StatsRef,** 2016. https://doi.org/10.1002/9781118445112.stat06405.pub2

LY, A; MARSMAN, M; VERHAGEN, J; GRASMAN, R. P. P. P; WAGENMAKERS, E.-J. A tutorial on Fisher information. **Journal of Mathematical Psychology**, 80, p. 40–55, 2017. https://doi.org/10.1016/j.jmp.2017.05.006

MACCANN, C; ROBERTS, R. D; New Paradigms for Assessing Emotional Intelligence: Theory and Data, **Emotion,** v. 8**,** n. 4, p. 540–551, 2008. https://doi.org/10.1037/a0012746

MANCINI, G.; BIOLCATI, R.; JOSEPH, D.; TROMBINI, E.; ANDREI, F. Emotional intelligence: Current research and future perspectives on mental health and individual differences. **Fronters of Psychology**, v. 13, Article 1049431, 2022.

MARTIN, A. J; LAZENDIC, G. Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. **Journal of Educational Psychology,** v. 110**,** n. 1, p. 27–45, 2018. https://doi.org/10.1037/edu0000205

MAYER, J.D.; CARUSO, D. R.; SALOVEY, P. The ability model of emotional intelligence: Principles and updates. **Emotion Review**, v. 8, n. 4, p. 290–300, 2016. https://doi.org/10.1177/1754073916639667

MAYER, J. D; SALOVEY, P. What is emotional intelligence? In P. Salovey & D. J. Sluyter (Eds.), **Emotional development and emotional intelligence: Educational implications** (pp. 3–31). Basic Books, 1997.

MAYER, J. D; SALOVEY, P; CARUSO, D. R; SITARENIOS, G. Measuring emotional intelligence with the MSCEIT V2.0. **Emotion,** v. 3**,** n. 1, p. 97–105, 2003. https://doi.org/10.1037/1528-3542.3.1.97

MURAKI, E. A generalized partial credit model: Application of an EM algorithm. **Applied Psychological Measurement,** v. 16**,** n. 2, p. 159–176. https://doi.org/10.1177/014662169201600206

MURPHY, D. L, DODD, B. G; VAUGHN, B. K. A Comparison of Item Selection Techniques for Testlets. **Applied Psychological Measurement,** v. 34**,** n. 6, p. 424–437, 2010. https://doi.org/10.1177/0146621609349804

OLIVEIRA, G. F; BARBOSA, G. A; SOUZA, L. E; COSTA, C. L; ARAUJO, R. C; GOUVEIA, V. V. Satisfação com a vida entre profissionais da saúde: correlatos demográficos e laborais, **Revista de Bioética**, v. 17**,** n. 2, p. 319–334, 2009.

OLIVEIRA, C. M. **Construção e busca de evidências de validade de um banco de itens de personalidade para testagem adaptativa desenvolvido a partir dos princípios do desenho universal**. Tese (Doutorado em Psicologia) Departamento de Psicologia, Universidade Federal da Santa Catarina. Repositório Institucional da UFSC, p. 191, 2017. https://repositorio.ufsc.br/bitstream/handle/123456789/187269/PPSI0766-T.pdf

PASSOS, M. F. D., & LAROS, J. Construção de uma escala reduzida de Cinco Grandes Fatores de personalidade, **Avaliação Psicológica,** v. 14**,** n. 1, p. 115–123, 2015. https://doi.org/10.15689/ap.2015.1401.13

PENFIELD, R. D. Applying Bayesian item selection approaches to adaptive tests using polytomous items. **Applied Measurement in Education,** v. 19, p. 1–20, 2006. https://doi.org/10.1207/s15324818ame1901_1

PERES, A. J. de S. Testagem adaptativa por computador (CAT): Aspectos Conceituais e um Panorama da Produção Brasileira. **Examen: Política, Gestão E Avaliação Da Educação**, v. 3**,** n. 3, p. 66–86, 2019. https://examen.emnuvens.com.br/rev/article/view/10

REEVE, B. B. Special Issues for Building Computerized-Adaptive Tests for Measuring Patient-Reported Outcomes: The National Institute of Health's Investment in New Technology**. Medical Care,** v. 44, n. 11 (Suppl 3), S198–S204, 2006. https://doi.org/10.1097/01.mlr.0000245146.77104.50

SANTOS, J. S. **Mensuração de habilidades cognitivas preditoras do desenvolvimento de leitura em crianças através de jogos educacionais** [Master's thesis, Universidade Federal de Campina Grande]. Repositório da Universidade Federal de Campina Grande, 2017.

SHANNON, C. E. A mathematical theory of communication. **The Bell System Technical Journal,** v. 27**,** n. 3, p. 379–423, 1948. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

SOUZA, V. V. **Adaptação de uma Medida de Inteligência Emocional para um Teste Adaptativo Computadorizado para o Contexto Brasileiro**. 2023. Dissertation (Doctoral Degree) – Universidade de Brasília, Brasília. Repositório Institucional da UnB.

VAN DER LINDEN, W. J. Bayesian item selection criteria for adaptive testing. **Psychometrika**, v. 63, p. 201–216, 1998. https://doi.org/10.1007/BF02294775

VAN DER LINDEN, W. J., & PASHLEY, P. J. Item Selection and Ability Estimation in Adaptive Testing. In W. J. van der Linden, G. A. W. Glas (Eds.), **Computerized Adaptive Testing: Theory and Practice**. Springer, 2000. https://doi.org/10.1007/0-306-47531-6_1

VEERKAMP, W. J. J., & BERGER, M. P. F. Some New Item Selection Criteria for Adaptive Testing**. Journal of Educational and Behavioral Statistics**, v. 22**,** n. 2, p. 203–223, 1997. https://doi.org/10.2307/1165378

VIEIRA-SANTOS, J., LIMA, D. C., SARTORI, R. M., SCHELINI, P. W., & MUNIZ, M. Inteligência emocional: revisão internacional da literatura. **Estudos Interdisciplinares em Psicologia**, v. 9**,** n. 2, p. 78–99, 2018. https://doi.org/10.5433/2236-6407.2018v9n2p78

WICKHAM, H. **ggplot2:** Elegant Graphics for Data Analysis**.** Second Edition. Springer, 2016.

WISE, S. L. (2018). Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating**. Education Inquiry,** v. 10*,* n. 1, p. 1–13, 2018. https://doi.org/10.1080/20004508.2018.1490127

Recebido em*: 20 de outubro 2023*

Aceito em: *24 de janeiro 2024*