https://doi.org/10.18222/eae.v0ix.7216

A COMPUTERIZED ADAPTIVE TEST OF PROVINHA BRASIL – LEITURA: RESULTS AND PERSPECTIVES

RODRIGO TRAVITZKI[I]
OCIMAR MUNHOZ ALAVARSE[II]
DOUGLAS DE RIZZO MENEGHETTI[III]
ÉRICA MARIA DE TOLEDO CATALANI[IV]
TRANSLATED BY: FERNANDO EFFORI[V]

[I] Universidade São Francisco (USF), Campinas-SP, Brazil; *r.travitzki@gmail.com*
[II] Universidade de São Paulo (USP), São Paulo-SP, Brazil; *ocimar@usp.br*
[III] Centro Universitário da Fundação Educacional Inaciana "Padre Sabóia de Medeiros" (Centro Universitário FEI), São Bernardo do Campo-SP, Brazil; *douglasrizzo@fei.edu.br*
[IV] Universidade de São Paulo (USP), São Paulo-SP, Brazil; *ericamtc@usp.br*
[V] Freelance translator, São Paulo-SP, Brazil; *feffori@gmail.com*

ABSTRACT

This article describes a Computerized Adaptive Test (CAT) of Provinha Brasil – Leitura,[1] based on Item Response Theory. We detail the operation and development of the algorithm. The CAT was administered by means of tablet computers to 1,983 students in the 1st and 2nd grades of primary education, in 15 schools of the Municipal Education System of São Paulo. Results confirm the quality of Provinha Brasil's items, of the work done in schools and, mainly, of the CAT. As to the management of test time, we found a positive association between proficiency and time, but only to a certain extent; students tend to take longer on the more difficult items; this tendency is stronger in more proficient students, thus confirming the hypothesis that they tend to manage test time better.

KEYWORDS COMPUTERIZED ADAPTIVE TESTING • COMPETENCY ASSESSMENT • PROVINHA BRASIL • ITEM RESPONSE THEORY.

[1] Provinha Brasil – Leitura is a Brazilian standardized reading assessment test for primary education students.

TESTE ADAPTATIVO INFORMATIZADO DA PROVINHA BRASIL – LEITURA: RESULTADOS E PERSPECTIVAS

RESUMO

Este artigo descreve um Teste Adaptativo Informatizado (TAI) da Provinha Brasil – Leitura, com base na Teoria da Resposta ao Item. Detalham-se o funcionamento e o desenvolvimento do algoritmo. O TAI foi aplicado com o uso de tablets a 1.983 alunos dos 1º e 2º anos do ensino fundamental, em 15 escolas da Rede Municipal de Ensino de São Paulo. Os resultados confirmam a qualidade dos itens da Provinha Brasil, do trabalho realizado nas escolas e, sobretudo, do TAI. Em relação à gestão do tempo de prova, conclui-se que há uma associação positiva entre proficiência e tempo, mas só até certo ponto; os alunos tendem a demorar mais nos itens mais difíceis; essa tendência é mais intensa nos alunos mais proficientes, confirmando a hipótese de que eles tendem a gerir melhor o tempo de prova.

PALAVRAS-CHAVE TESTE ADAPTATIVO INFORMATIZADO • AVALIAÇÃO DE COMPETÊNCIA • PROVINHA BRASIL • TEORIA DA RESPOSTA AO ITEM.

TEST ADAPTATIVO INFORMATIZADO DEL PROVINHA BRASIL – LEITURA: RESULTADOS Y PERSPECTIVAS

RESUMEN

Este artículo describe un Test Adaptativo Informatizado (TAI) del la Provinha Brasil– Leitura, con base en la Teoría de la Respuesta al Ítem. Se detalla el funcionamiento y el desarrollo del algoritmo. El TAI fue aplicado con el uso de tablets a 1.983 alumnos del primero y segundo año de la enseñanza primaria, en 15 Escuelas de la Red Municipal de Enseñanza en San Pablo, Brasil. Los resultados confirman la calidad de los ítems del la Provinha Brasil, del trabajo realizado en las escuelas y, especialmente, del TAI. En relación con la administración del tiempo del examen, se concluye que hay una asociación positiva entre proficiencia y tiempo, pero solo hasta cierto punto; los alumnos tienden a demorarse más en los ítems más difíciles; esta tendencia es más intensa en los alumnos más proficientes, confirmando la hipótesis de que ellos tienden a administrar mejor el tiempo del examen.

PALABRAS CLAVE TEST ADAPTATIVO INFORMATIZADO • EVALUACIÓN DE COMPETENCIA • PROVINHA BRASIL • TEORÍA DE LA RESPUESTA AL ÍTEM.

INTRODUCTION[2]

Standardized tests, mainly due to their use in external assessments, have become increasingly present in public schools and, one may suppose, are gradually entering the Brazilian school culture as an expression of educational policies that use these tests through the several uses of their results. However, if this type of policy is to contribute to improving the quality of education, several challenges must be faced. Some of them concern the interpretation of results and how they are used in strategies for managing education systems, and even more in the classroom on an everyday basis. Other challenges refer to logistical difficulties related to the security and management of large amounts of paper. A third type of challenge concerns the technical quality of tests as measurement instruments. Computerized Adaptive Testing (CAT) is a learning assessment technology that can help overcome these challenges, especially the latter two types.

The idea of a test that adapts to the proficiency of each individual dates back to the 1970's, although it has not yet become popular, and it is incipient in Brazil. In this type of test, each individual responds to a set of items, selected according to their mastery – their proficiency – of what is being evaluated in the course of the test, such that the test is only fully defined when the respondent has finished it. Therefore, with CAT, a test may vary from one respondent to another due to their differences in proficiency. At the beginning of the test, each respondent is presented with an item, which can, for example, be randomly selected or chosen according to a preset condition. Then, as individuals are taking the test, it adapts to each individual, depending on their performance (represented by their correct and incorrect answers), mainly through the selection of more difficult or easier items; it aims both at optimizing the accuracy in estimating each individual's proficiency and at reducing administration time, in contrast to tests in paper or presented on a computer, considered linear, which, from the beginning, are ready with a certain number of items, even if presented sequentially. This is the essence of an adaptive test, which can have several variations and complexity levels. When this type of test is administered in a digital format, by means of computerized processes, it is called a Computerized Adaptive Test.

Compared to a conventional linear test, CAT has some advantages, for example: 1) it provides equalized measurement accuracy for different levels of proficiency;[3]

[2] Part of this study was presented at the 2018 Teaching, Learning and E-learning (IAC-TLEI) congress, in Vienna, Austria. We are grateful to the various education professionals who collaborated for the pilot project, the Technical Evaluation Center of the Municipal Education Department of São Paulo, the Faculty of Education of the University of São Paulo (USP), the National Institute for Educational Studies and Research Anísio Teixeira (INEP), and to the United Nations Educational, Scientific and Cultural Organization (UNESCO).

[3] Considering that, in conventional tests, there is greater inaccuracy in proficiency estimates at the lower and upper extremes of the scale.

2) it allows maintaining the test's accuracy while significantly reducing the number of items; and 3) it allows increasing the test's accuracy in case the number of items is maintained (BARRADA, 2012). Of course, the feasibility of CAT depends, like any assessment methodology, on an item pool with a great variability and density of difficulty in relation to the proficiency scale.

In an adaptive test, after the answer given to the first item, items are selected according to the previous answers. Thus, for each answer, some information is added about the respondent's proficiency measure, a process by which the uncertainty (error) in the proficiency estimate is reduced. In other words, the measurement instrument – the test – becomes more accurate with each answer. In addition, real-time selection of items allows collecting more qualified information about students' proficiency at the scale's less explored points, such as the upper and lower extremes.

To better understand this idea, we may simply imagine that a student has given correct answers to six easy items; in this case, there would be no need for the student to answer more easy items, since one more correct answer to an easy item would add little information to the estimate of his proficiency as it is reasonably certain that the student has sufficient proficiency to give right answers to items in that part of the scale. Thus, a new easy item will likely not change the proficiency estimate, nor will it reduce the estimate's uncertainty; it is necessary to increase the difficulty of the next items to be presented to a point where the termination criterion is met. This is one of the great advantages of adaptive tests: optimizing the collection of information about the student's proficiency by avoiding unnecessary measuring through uninformative items, because

> Within a CAT format, item selection and ability estimation proceed hand in hand. Efficiencies in ability estimation are heavily related to the selection of appropriate items for an individual. In a circular fashion, the appropriateness of items for an individual depends in large part on the quality of interim ability estimates. (LINDEN; GLAS, 2010, p. 4)

In the CAT of *Provinha Brasil – Leitura*, which is object of this article, a pool with 39 items provided by the National Institute for Educational Studies and Research Anísio Teixeira (INEP) was used, and we tried to build a test with fewer items and greater accuracy for each student, on average, than the original linear test administered in paper and comprising 20 multiple-choice items. Furthermore, in addition to decreasing the number of items and increasing accuracy, we tried to solve a problem that was not addressed in the literature, i.e., that, in a formative assessment, the test termination criterion should take into account not only the uncertainty regarding the estimated proficiency, but also the uncertainty regarding the proficiency level attributed to the examinee, i.e., his classification in one of the segments of the proficiency scale. This is mainly because each proficiency level corresponds to a pedagogical interpretation, and the idea behind a test is that its result be used by teachers in their daily activities, when it is more important to know the description of what the student is able to do at that level than to receive a number – the estimated proficiency – that would only allow comparing students' performances. Based on this more qualitative information, teachers can plan didactic activities to reach higher levels. Additionally, the assessment – a judgment on the student's performance – is made according to criteria that relate performance and school grade to these levels of pedagogical interpretation.

If, on the one hand, the science of adaptive testing is still evolving and revealing its great potential, on the other, there are also considerable challenges regarding the complexity of statistical procedures compared to that complexity in linear tests (LINDEN; GLAS, 2010). For example, there has to be one or more criteria for the selection of items, and choosing a criterion may depend on factors such as the goals of the test, item pool size and the distribution of items' difficulties. It is also important to consider issues related to item pool security and sustainability, which can be optimized with different techniques to control items' exposure rate (BARRADA, 2010). In addition, CAT needs one or more termination criteria. The commonly used criteria are: 1) reaching a preset number of items; 2) reaching a minimum level of uncertainty in the proficiency estimate; and 3) reaching a minimum threshold of information that a new item would add to the proficiency estimate (BARRADA, 2012).

As previously considered, item pool quality and size are an important and challenging aspect, the solution of which depends on the test's goals and on specific initiatives for creating items. For Barrada (2012), there are four general goals for CAT which can be given greater or lesser importance: 1) reliability of the proficiency estimate; 2) item pool security; 3) content restrictions; and 4) item pool maintenance. Some of these goals oppose one another, such as (1) in relation to (2) and (4). Indeed, there is a trade-off between measurement accuracy and item pool security,[4] which can be minimized with other item selection methods, to the detriment of selecting the most informative item at each point in the test (GEORGIADOU; TRIANTAFILLOU; ECONOMIDES, 2007).

The CAT described in this article is a computerized adaptive version of Provinha Brasil, a standardized instrument provided by the Brazilian Ministry

[4] Item pool security refers, for example, to the exposure rate of items. The more an item is presented to a population, the less informative it tends to be, especially in high-stakes testing.

of Education (MEC) which was created in 2007 (BRASIL, 2007a) to assess reading proficiency, and its results are expressed on a scale of five levels numbered from 1 to 5. The CAT described here was administered in 15 primary education schools of the Municipal Education System of São Paulo. The article has a first section that briefly describes Provinha Brasil – Leitura, then it focuses on the operation of the CAT's algorithm, which is the component responsible for the "adaptive" part of the test. The third section addresses the computational simulations used to build the algorithm and perform the initial adjustment of some parameters. The subsequent section presents the results from the CAT's administration, which confirmed the validity of the instrument. The fifth section deals with the CAT's experimental administration so as to analyze, among other results, a more general question based on the data collected which regards the way students manage test time. The conclusions summarize the results of the CAT of Provinha Brasil – Leitura and some indications for further research.

PROVINHA BRASIL – LEITURA

Literacy acquisition is especially important in the initial years of primary education, a period when teachers dedicate much of their teaching to develop children's reading and writing skills (SOARES, 2016). National indicators have shown unsatisfactory results in the development of these skills for the entire Brazilian population in school age. In this context, the CAT's construction process turned its focus to Provinha Brasil – Leitura,[5] which was developed by the INEP for students in the second grade of primary education, based on the Education Development Plan (PDE) (BRASIL, 2007b), in line with international organizations' recommendations for the "literacy decade" (2003 to 2012) (GONTIJO, 2012).

The construction of a computerized adaptive version of Provinha Brasil – Leitura was the result of a project developed by researchers at the Study and Research Group on Educational Assessment (GEPAVE) linked to the Faculty of Education of the University of São Paulo (FEUSP) in partnership, during the project's initial stage, with the INEP, and, throughout its development, with the Municipal Education Department of São Paulo (SME-SP), by means of its Technical Evaluation Center (NTA), involving the department's central management, regional managers (school supervisors), school principals and pedagogical coordinators, teachers and students.[6] Giving due importance to reading in the development of a consequent pedagogical project requires, among other elements, dedicating attention to

[5] Provinha Brasil incorporated Mathematics competency as of 2011, but the CAT project focused on reading, considering its importance in the schooling process.

[6] For more details on the project, see Alavarse et al. (2018) and Catalani (2019).

assessment procedures (SOARES, 2016), considering the formative approach to the use of their results, and Provinha Brasil – Leitura is an instrument that was built with that approach. It is based on a Guiding Matrix that considers the skills related to basic literacy acquisition, understood as development of the understanding of the rules pertaining to the alphabetical writing system, and functional literacy acquisition, understood as apprehension of possibilities of social uses and functions of written language.[7] Within the scope of Provinha Brasil, basic literacy acquisition and functional literacy acquisition are approached as complementary and parallel processes, and the items that make up the tests seek to cover them in the test's spectrum of difficulty. It is worth noting that the items are pre-tested across the country, and that the aforementioned matrix is based on several MEC documents dealing with the training of teachers for the initial years of primary education. The aim of all these measures is to guarantee both the reliability of the test's results and its validity for purposes of use of such results, a condition for them to be fully integrated into the teaching process in the initial years of primary education.

Administered since 2008 and discontinued in 2016, Provinha Brasil consists of two tests, the first of which for administration in March (beginning of the school year), and the second, in October (end of the school year), both containing 20 multiple choice items created according to a matrix of guidelines for reading. The items were pre-tested according to statistical standards, calibrated by INEP specialists and delivered to teachers across the country, along with instructions on administration, "grading" – in fact, response processing – and interpretation of the results.

These initial results, formed by students' scores in the tests, are later expressed in a reading proficiency scale with five proficiency levels, including a pedagogical interpretation for each level, as well as teaching suggestions for students' progress. The quality of this process is founded on the pre-testing of items so as to allow this association between scores and proficiency, which was confirmed, in the case of this CAT, in discussions with more than 100 teachers at the participant schools.

In the paper version of Provinha Brasil, the teachers themselves administer the test, tabulate the responses and interpret the results, with all test items being delivered to the teachers and managers at each administration. This procedure, according to teachers' opinions, allows them to better understand the results, as can be seen in a case study in the municipality of Camaragibe, which described

[7] Note of the translator: In Brazil, the concept of literacy acquisition is often divided into the two concepts presented in this passage. The former, i.e., basic literacy acquisition, is designated in the Brazilian literature by the Portuguese alfabetização, whereas the latter, i.e., functional literacy acquisition, is designated by the Portuguese letramento inicial. Those are the terms are used in the original Portuguese version of this article.

an interesting appropriation of the instrument and highlighted the importance of constant work by teachers and managers for the good results reported (MORAIS; LEAL; ALBUQUERQUE, 2009). Although it is not the focus of this study, it is worth noting the presence of dissent in relation to the concept of reading implicit in the guiding matrix and in Provinha Brasil's test structure, as found in Gontijo (2012); however, as pointed out, Provinha Brasil's potential is acknowledged with regard to the support it provides to teaching, since the result of its administration is not limited to a number representing the score of correct answers or even the student's level of proficiency, because in the administration material of each test, each level is accompanied by a description of what the student is able to do in terms of reading skills, as well as suggestions for teaching intervention so that the student can progress to a higher level.

It is important to note that, in the item pool used, each item was associated with a descriptor in the Reading Matrix, i.e., each element in this matrix that, as a whole, describes the "reading" construct, for which the proficiency is being assessed, and which, in the documents of Provinha Brasil, is considered an

> [...] activity that depends on individual processing, but is part of a social context and involves [...] capacities related to deciphering, understanding and producing meaning. This approach to reading, therefore, encompasses from skills necessary for the process of basic literacy acquisition to those that enable the student to actively participate in literate social practices, i.e., those that contribute in his functional literacy acquisition process. This implies that the student develop, among other skills, those consisting in reading words and sentences, finding explicit information in sentences or texts, recognizing the subject of a text, recognizing the purposes of texts, making inferences and establishing relationships between parts of a text.[8] (BRASIL, 2016, p. 9, own translation)

In these terms, not only does Provinha Brasil constitute material under greater control by teachers for purposes of administration and treatment of responses, but it also allows them full access to its fundamentals and the scope of its interpreted results, in addition to providing alternatives for didactic developments, which

[8] In the original: *"atividade que depende de processamento individual, mas se insere num contexto social e envolve [...] capacidades relativas à decifração, à compreensão e à produção de sentido. A abordagem dada à leitura abrange, portanto, desde capacidades necessárias ao processo de alfabetização até aquelas que habilitam o(a) estudante à participação ativa nas práticas sociais letradas, ou seja, aquelas que contribuem para o seu letramento. Isso implica que o(a) estudante desenvolva, entre outras habilidades, as de ler palavras e frases, localizar informações explícitas em frases ou textos, reconhecer o assunto de um texto, reconhecer finalidades dos textos, realizar inferências e estabelecer relações entre partes do texto".*

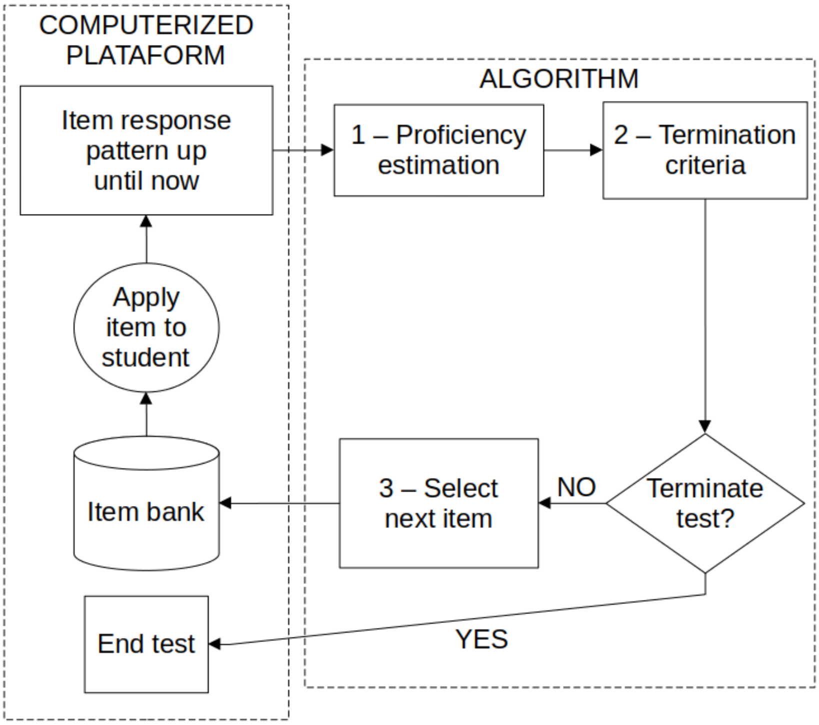
promotes debate in schools, since it is more transparent than other external learning assessments.

OPERATION OF THE CAT'S ALGORITHM

There are several challenges to making a learning assessment test feasible. In the case of a CAT, in addition to the traditional ones, which concern item quality and the validity of the test as a whole, two others are added: being computerized and being adaptive so as to provide benefits in terms of logistics and proficiency measurement accuracy. To face both challenges, we chose to build two relatively independent modules: a user-friendly interface (created with web-based technology using Java as the main programming language) and a real-time statistical and psychometric processing system (developed in R language). The former module allows the test to be computerized, whereas the latter makes it adaptive. The interface will be referred to in this study as "computerized platform", whereas the statistical processing will be called "algorithm".

The computerized platform is responsible for displaying items to respondents and for automatically capturing responses. It works online and was accessed via an intranet for all participant schools. The items were presented by means of tablets connected to the schools' Wi-Fi networks. The reading of the items' statements for the students, which in Provinha's original version must be performed by the administrator teacher, was performed via this platform, i.e., it was individually provided to the students, since they were using tablets fitted with headphones and Google Speech API. Figure 1 represents, in a simplified way, the components of the CAT of Provinha Brasil – Leitura and their interrelationships.

FIGURE 1 – Simplified representation of the components of the CAT of Provinha Brasil – Leitura and their interrelationships



Source: Prepared by the authors.

The algorithm is theoretically based on Item Response Theory (IRT), described, among other authors, by Baker (2001), and its goal is to provide an adaptive dynamic to the computerized platform. More specifically, we sought to optimize measurement accuracy and minimize the number of test items, avoiding losses in the instrument's validity and in computational processing time.

The algorithm comprises three components, each characterized by a sequence of actions, as described below.

1) Proficiency estimation:
   a) receives as input from the platform the responses of each student – thus setting a pattern of correct answers – and the a priori distribution of proficiency, in addition to the item pool's parameters;
   b) estimates the proficiency and standard error using a Bayesian method that incorporates the a priori distribution.
2) Test termination criterion:
   a) checks whether the test has reached the maximum limit of items;
   b) checks whether the standard error is below the maximum limit defined;
   c) checks whether the proficiency level has been reliably identified;
   d) if at least one of the three conditions above is true, and if the test has exceeded the minimum limit of items, then it sends an output to the platform to terminate the test; otherwise, it proceeds to the third step.

3) Selection of the test's next item:
    a) identifies the least represented matrix descriptor among the items administered to the student so far;
    b) searches for the items for this descriptor in the item pool;
    c) searches for the most informative item on this subset, taking into account the proficiency estimated in step 1;
    d) returns the selected item as an output to the platform (along with the estimated proficiency and standard error, so that the platform can include them as information the next time the algorithm is activated, after the selected item has been responded).

Proficiency estimation is performed based on the expected a posteriori (EAP) (BOCK; MISLEVY, 1982) distribution with a 21-point quadrature. The criteria for selecting items include: 1) Fisher's Maximum Information (MFI) (BARRADA, 2010); and 2) the balanced selection of items from each core area in the matrix.

The balanced selection of items from the guiding matrix descriptors aims to avoid the loss of test validity due to the selection of items performed by the CAT. In fact, this is an important precaution to avoid a side effect from the use of Fisher's maximum information as an item selection criterion. The algorithm always maintains a similar proportion of items from each core area to ensure that the adaptive test represents the desired matrix. However, although it was implemented in the algorithm, the balanced selection of items was not applied to this pilot phase due to the small number of items in the pool.

To determine the end of the test, the algorithm uses a mixed criterion, taking into account three rules:

a) number of test items (minimum of eight and maximum of 20 items);
b) uncertainty limit (standard error below 35 points);
c) degree of confidence in determining the proficiency level in the scale of Provinha Brasil – Leitura (85% confidence, according to our simulations).

The first two criteria are widely used in adaptive testing (BARRADA, 2012). The third criterion (c), developed for the CAT of Provinha Brasil – Leitura, constitutes a change in the termination criterion used in assessments in order to classify the respondents. The classification is used in summative assessment contexts, in which a yes/no decision must be made, which is more common in exams and certifications (BABCOCK; WEISS, 2012; WEISS; KINGSBURY, 1984). This is apparently a significant contribution of this project to the state of the art of CAT.

SIMULATIONS OF THE CAT OF PROVINHA BRASIL – LEITURA

This section presents the technologies and simulation methodology used to develop the CAT of Provinha Brasil – Leitura.

Software and Hardware

The algorithm was written in R, a free and open source programming language that is specialized in statistics. The simulations were also carried out in this language. For both purposes, the following packages were tested: *catR* (MAGIS; RAÎCHE, 2012), *PP* (REIF, 2019) and *irtoys* (PARTCHEV, 2016). The simulations were initially carried out in 2016 and then repeated in February 2018 with new versions of the packages in order to preserve the quality of results. The versions of the packages used in the 2018 simulations, presented in this study, were: *irtoys 0.2.0*; *PP 0.6.1*; and *catR 3.13*.

The computer used to carry out the simulations was a notebook with an i7 4-core 2.50 GHz processor and 8 Gb of memory, with Linux Mint operating system. No parallel processing or acceleration via GPU was used.

Simulations

To develop the algorithm, five item selection methods (in addition to random selection) and seven proficiency estimation methods were tested via simulation. The methods were tested for accuracy and speed. In addition, the simulations allowed the adjustment of two parameters in the algorithm: maximum standard error and confidence interval critical value.

The simulations are based on the two-parameter logistic function of IRT (ANDRADE; TAVARES; VALLE, 2000), which describes the probability that an individual with known proficiency answers correctly an item with known parameters. Although different pools were tested, the results described here refer to the item pool provided by INEP, which is formed by 39 items with two defined parameters (difficulty and discrimination).

For each situation, 1,000 simulations were carried out, each with 1,000 participants (with normal distribution of proficiency, with a mean of 500 and standard deviation of 100) responding to 20 of the 39 items from the original INEP pool.

To estimate proficiency, four methods were compared, one of which was tested in several packages, thus totaling seven methods. Two methods are based on the likelihood principle: the search for maximum likelihood (LORD, 1980) and weighted likelihood (WARM, 1989). The other two methods use Bayesian statistics: expected a posteriori (EAP) (BOCK; MISLEVY, 1982) distribution and the modal estimator (BIRNBAUM, 1969).

Below, we list the seven compared methods from the three packages.

1) ML: maximum likelihood (*catR* package);
2) WL: weighted likelihood (from the *catR* package);
3) BM: Bayesian modal estimator (from the *catR* package);
4) EAP: EAP method (thetaEst function from the *catR* package);
5) eapC: EAP method (eapEst function from the *catR* package);
6) eapI: EAP method (from the *irtoys* package);
7) eapP: EAP method (from the *PP* package).

The item selection methods tested from the *catR* package (MAGIS; RAÎCHE, 2012) were the following:

a) *random*: random selection of pool items;
b) MFI: it selects the item with the highest information for the proficiency estimated so far, based on the item's information function (ANDRADE; TAVARES; VALLE, 2000);
c) bOpt (Urry's rule): it selects the item with the closest difficulty level to the proficiency estimated so far;
d) thOpt (stratification by Maximum Information): an adaptation of the FMI method in order to increase item pool security;
e) The progressive method (REVUELTA J.; PONSODA, 1998): the item is selected according to two elements, a Maximum Information-related one and a random one. As the test progresses, the random element loses relevance. That helps increase item pool security; and
f) The proportional method (BARRADA, 2010): the item is selected according to probabilities related to Fisher's Information, also in order to increase item pool security.

Test Termination Criteria

The main goal in designing the termination criterion was to provide a test with fewer items and greater accuracy than a similar though not adaptive test. To that end, three criteria were simultaneously considered.

First, we selected beforehand a maximum limit of 20 items and a minimum of 8 items. The minimum limit guarantees the administration of at least four items from each of Provinha's two core areas.[9] The maximum limit, on the other hand, guarantees the termination of the test at the levels of the paper version, even if the other criteria are not met.

Secondly, we defined a maximum error in the proficiency estimate, which is measured by the standard error. This error starts high and decreases as the test

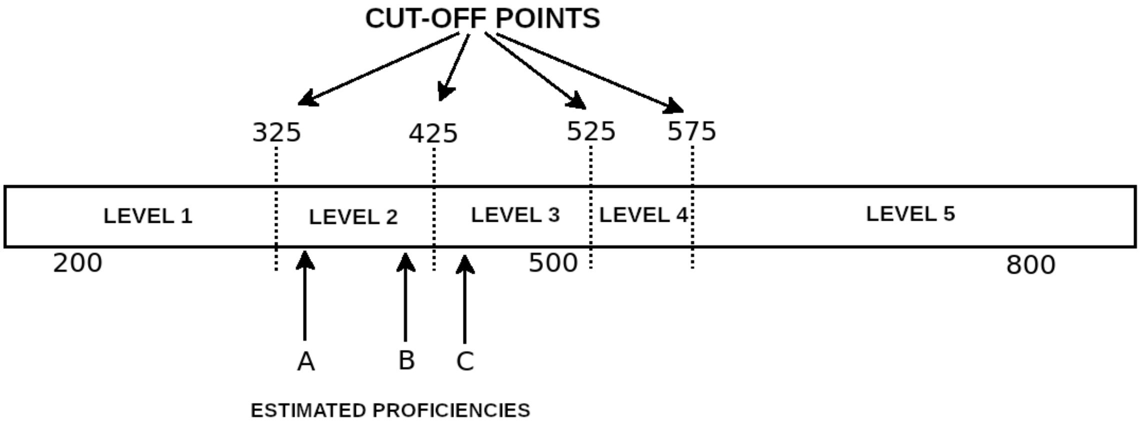
[9] The two core areas used in Provinha Brasil – Leitura are: 1) appropriation of the writing system; and 2) reading. However, due to the size of the item pool available, we chose not to use this criterion in the pilot.

progresses and more items are answered. The termination criterion consists in determining a maximum limit allowed for the standard error, according to the available items, the target population and the goals. And, finally, the criterion of "proficiency level reliability", which is described below.

Proficiency Level Reliability

This criterion seeks to optimize the size of the test while ensuring the student's correct allocation to the proficiency level. It determines the end of the test when the proficiency and its confidence interval are entirely contained in just one of the five proficiency levels defined for *Provinha Brasil* – Leitura (Figure 2). It is worth mentioning that the points that divide the Provinha Brasil's scale into five levels, also called cut-off points, resulted from a psychometric (anchoring) process associated with the pedagogical analysis of the items by specialists and educators. It is the description of the levels – resulting from this process – that allows interpreting the student's grade (proficiency) based on Item Response Theory. Although this method was later identified in the literature as a termination method for classification purposes, its use for formative assessment purposes was not found in the literature on adaptive testing.

FIGURE 2 – Cut-off points and proficiency levels in the scale of Provinha Brazil – Leitura



Source: Prepared by the authors.

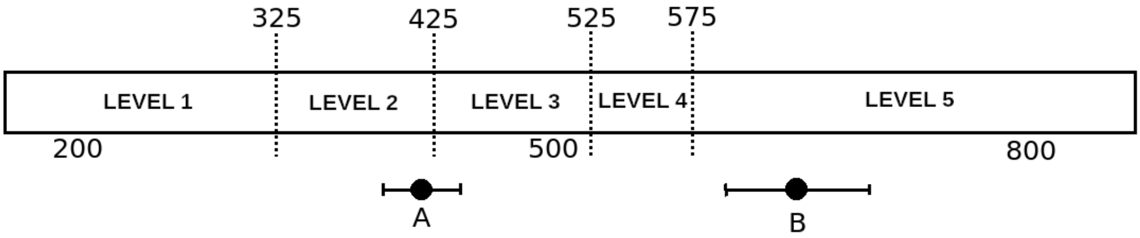
By adding this test termination criterion, the CAT of Provinha Brasil can be terminated even if the error in estimating a student's proficiency is relatively high, provided that the interval containing the proficiency is fully contained in one of the five levels of Provinha Brasil – Leitura. After all, Provinha is not a selection or certification test in which proficiency accuracy is more relevant. The most important thing is to know whether the student's proficiency level has been properly identified, thus prioritizing the pedagogical diagnosis and, consequently, the intervention to improve literacy. In other words, this criterion contributes to interpretation and assessment, which go beyond simple measurement.

By way of example, one can see in Figure 2 that the proficiency measured for student B is closer to the proficiency measured for student C than to that measured for student A. However, taking into account the cut-off points defined by the pedagogical interpretation of the scale, students A and B are at the same level in the scale, while student C is at the next level. In pedagogical terms, this means that students A and B supposedly demonstrated a mastery of skills that requires similar interventions, while student C demonstrated a broader mastery that requires other interventions.

It is worth highlighting that there is always some uncertainty about the estimation of a student's proficiency, regardless of the method used. That uncertainty depends mainly on the number of items answered and also on the closeness between the item's difficulty and the respondent's proficiency. Assuming a normal distribution of proficiency, the estimate's reliability (its confidence interval) can be determined based on the standard error. The confidence interval is defined by a minimum and a maximum limit, and it can be obtained with different degrees of confidence by multiplying the standard error by the critical value, which in turn depends on the desired degree of confidence. For the algorithm, a confidence level of 85% was defined, which corresponds to a critical value of 1.44 (FERREIRA, 2005). This means that when the test is terminated by this criterion, there is an 85% probability that the student's true proficiency is within the level identified by the test, according to Item Response Theory.

Figure 3 illustrates the usefulness of the termination criterion by proficiency level reliability as a complement to the maximum standard error criterion. Student B has a greater error (confidence interval) than student A, but his proficiency level has already been soundly estimated (level 5) after he answered five items correctly in an adaptive test. Student A, in turn, responded to 11 items in this adaptive test, but has not yet been reliably classified, and may fit in levels 2 or 3. After all, the estimate's error does not depend only on the number of items presented, but also on the pattern of correct answers for each student and the parameters of the items answered. What is more, the number and position of the cut-off points strongly interfere with this criterion. Indeed, student B was benefitted by the addition of this third test termination criterion: finishing more quickly without loss in test accuracy, since the practical purpose of Provinha Brasil – Leitura is to provide a reliable measure of the student's proficiency so that teachers, considering the five proficiency levels, can make pedagogical decisions based on more reliable information. It is worth noting, without going into the merit of it, that the INEP defined as desirable for each student to be at least at level 4 by the end of the 2nd grade.

FIGURE 3 – Representation of the confidence intervals for two estimated proficiencies



Source: Prepared by the authors.

SIMULATION RESULTS

This section presents the results obtained via simulation in the stage of building the algorithm and adjusting the basic parameters.

Item Selection and Proficiency Estimation Methods

The item selection and proficiency estimation methods were tested for accuracy and processing speed. With regard to item selection, all methods proved fast enough. T-test revealed that all criteria had a smaller error than random selection, although the differences between the methods were not significant ($p < 0.05$) in most simulations. Therefore, taking into account the specialized literature, we chose to include in the algorithm the Fisher's Maximum Information as the selection method. However, should the CAT of Provinha Brasil – or any other CAT – consolidate as a public policy, this technical choice would have to be revised, since it does not take into account the security or sustainability of the item pool. The methods known as progressive or proportional could be more appropriate in that case.

With regard to the proficiency estimation methods, no significant difference ($p < 0.05$) was found in the accuracy of the methods with small error (BM, EAP, eapC, eapI, WL) for a 20-item test and a population with a mean of 500. However, it is worth noting that in populations with a different mean than expected (600 or 400, for example), WL proved more accurate than the other methods, which is due to its lower dependence on an a priori population estimate.

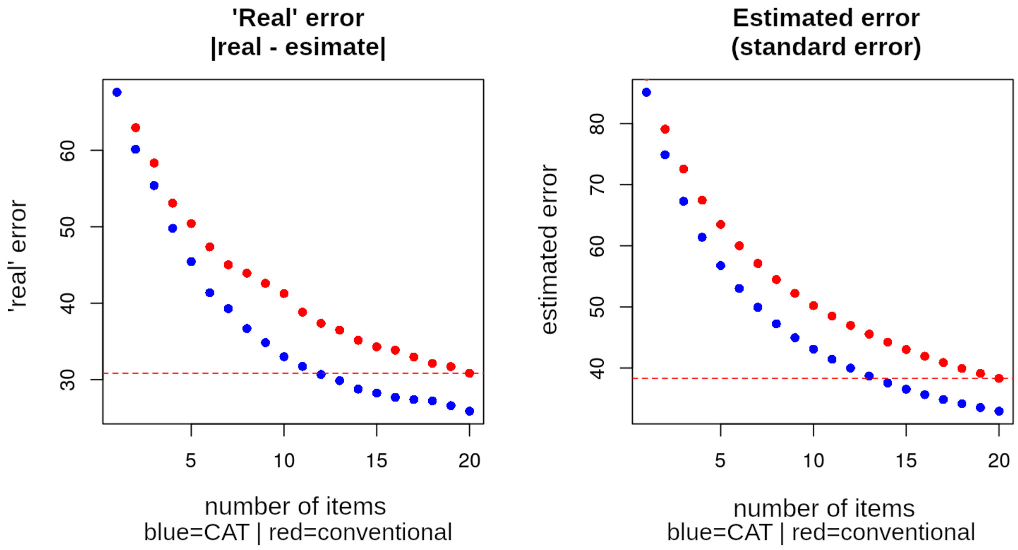
Also, the EAP method from the *PP package* (eapP) had a much greater error than the others. This illustrates another key role of simulations for the development of algorithms, which is to prevent the use of packages of dubious quality and with inconsistent results. Such caution is especially important when working with free software, but it is still necessary with proprietary software.

In sum, taking into account both accuracy and speed, the proficiency estimation method chosen was the EAP from the *irtoys* package. For next item selection, FMI was chosen, although this choice would have to be revised should it become necessary to preserve the item pool's security.

Test Termination Criterion: Maximum Standard Error

To determine the maximum standard error to be accepted by the CAT, we considered the goal of creating a test that is, on average, more accurate and smaller in size than a conventional, non-adaptive one. Figure 4 shows the estimation errors in conventional and adaptive tests according to the simulations performed. The estimated error is the one that can be obtained by the CAT algorithm for each new pattern of the examinee's responses. The "real" error, on the other hand, cannot be obtained by the algorithm. We only know it in the context of simulation.

FIGURE 4 – Estimation errors in conventional (normal) and adaptive (CAT) tests of different sizes up to 20 items, according to the simulations performed



Source: Prepared by the authors.
Note: The red dotted line indicates the error for the normal, 20-item test.

In the two graphs in Figure 4, the red dotted line marks the error reached by the conventional test after 20 items. This corresponds to a standard error of 38 and a difference in proficiencies of 31. Indeed, the graphs show that an adaptive test with 12-13 items tends to yield estimates as accurate as those of a conventional, 20-item test. Based on this observation, it is possible to delimit a midpoint, i.e., a standard error corresponding to a conventional test with 15-19 items. That adjustment depends on the balance desired between accuracy and test size in each assessment situation.

It is worth highlighting that in the two error detection methods there was this similarity, from 12 to 14 items, thus confirming the quality of the estimate obtained through the software used. Had there been no such similarity, the algorithm used would have no guarantee of corresponding to reality.

Finally, in order to provide a smaller, more accurate test, the maximum limit chosen for the standard error was 35 points in the *Provinha Brasil*'s scale, which would correspond to a 16-item adaptive test. Later, the results from the test's administration confirmed this prediction.

Proficiency Level Reliability

To define the proficiency levels, we used the cut-off points of Provinha Brasil – Leitura (Figure 2). Wider or narrower confidence intervals can be obtained for the same data set, depending on the confidence level desired. There is a critical value that corresponds to each confidence level. In order to adjust the best critical value to *Provinha Brasil*'s algorithm, we tested four confidence levels.

Table 1 shows, as expected, that the higher the confidence level desired, the less tests will be terminated according to the criterion of proficiency level reliability. It is important to know, however, how many tests were correctly terminated (comparing the "real" and the estimated proficiency levels) in each case. We identified how many of the tests terminated by this criterion would have been successful in determining the student's proficiency level.

TABLE 1 –Tests terminated by the criterion of proficiency level reliability according to confidence level in 1,000 simulations

CONFIDENCE LEVEL	CRITICAL VALUE	TESTS TERMINATED	CORRECT TERMINATIONS	CORRECTNESS RATE (%)
80%	1.28	354	297	83.9
85%	1.44	242	207	85.5
90%	1.645	114	109	95.6
95%	1.96	62	62	100.0

Source: Prepared by the authors.

These results show that, in Provinha Brasil – Leitura's scale, there is reasonable correspondence between the confidence level defined by the critical value and the confidence level of the proficiency level reliability criterion. In the algorithm, we set the critical value to 1.44.

RESULTS OF THE CAT'S EXPERIMENTAL ADMINISTRATION

We conducted 1,983 administrations of the CAT of Provinha Brasil – Leitura, with 823 students in the 1st grade and with 1,160 students in the 2nd grade of primary education, distributed in 80 classes in 15 primary education schools of the Municipal Education System of São Paulo.

TABLE 2 – Descriptive statistics for the experimental administration of the CAT of Provinha Brasil – Leitura to students in the 1st and 2nd grades of primary education of the Municipal Education System of São Paulo

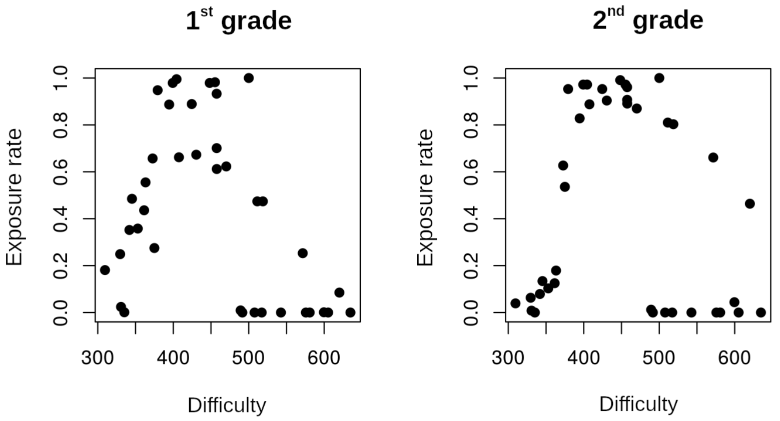
STATISTICS		PROFICIENCY	CORRECT ANSWERS	ITEMS	TEST SOLVING TIME	ADMINISTRATIONS
Total	Mean	462.72	0.55	17.33	11.52	1,983
	Std. Error	1.84	0.00	0.05	0.12	
	Std. Deviation	81.93	0.20	2.18	5.53	
1st grade	Mean	416.83	0.44	16.74	10.88	823
	Std. Error	2.14	0.01	0.06	0.20	
	Std. Deviation	61.41	0.15	1.60	5.75	
2nd grade	Mean	495.28	0.63	17.75	11.98	1,160
	Std. Error	2.32	0.01	0.07	0.16	
	Std. Deviation	79.04	0.19	2.43	5.32	

Source: Prepared by the authors.

Table 2 describes four important aspects of the CAT's administration. The average proficiency estimated for the 2nd grade (495.28 points) is of adequate magnitude, considering that Provinha Brasil's scale (designed for the 2nd grade) has a mean of 500 and a standard deviation of 100. In turn, the standard deviation (79.04) was below 100, which reflects a smaller variance of the population analyzed compared to the population for which Provinha Brasil was designed, i.e., students from all over Brazil. Also, the mean for the 1st grade was lower than that for the scale, which was also to be expected, since the items were designed and calibrated for the 2nd grade. These results confirm, therefore, not only the quality of the algorithm used in the CAT, but also the quality of the items designed by the INEP.

With regard to item exposure rate, the CAT did use some items more than others. In all, only 31 items were administered to the students. Considering that the pool has 39 items, 79.5% of the pool were put to use. Figure 5 aims to relate exposure rate and item difficulty. Again, the empirical results corroborate those expected via simulation: in addition to a higher rate for items of average difficulty, the figure also shows that, for the 2nd grade, there was a greater use of the more difficult items and a smaller use of easier items. It also shows that the vast majority of unused items is situated in the scale's upper portion.

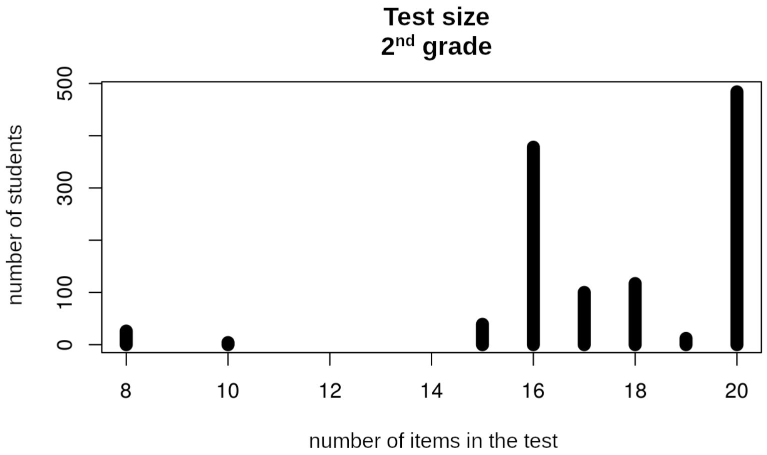
FIGURE 5 – Relationship between exposure rate and item difficulty for students in the 1st and 2nd grades of primary education



Source: Prepared by the authors.

With regard to the number of items presented to students, a good part of the tests were terminated with 16 items (Figure 6), as expected by the simulations and by the maximum limit of 35 points that we chose for the standard error of proficiency estimate. In addition, some tests were terminated with ten or fewer items, which was due to the other test termination criterion, i.e., proficiency level reliability. Thus, the results confirm the importance of this criterion as a complement to the criterion of maximum standard error. On the other hand, the great number of tests with 20 items was not predicted by the simulations, thus revealing limitations in the model used. Considering that the algorithm aims to keep a certain minimum degree of uncertainty, there may be real effects that were not predicted by the unidimensional two-parameter model of IRT that we used as the basis of our simulations.

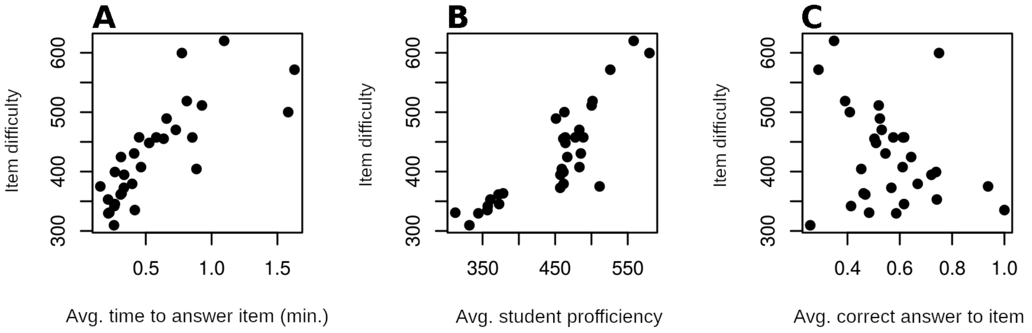
FIGURE 6 – Number of administrations of the CAT of Provinha Brasil – Leitura to students in the 2nd grade of primary education according to test size



Source: Prepared by the authors.

As expected, the more difficult items were presented to students with greater proficiency, as shown by the tendency in Figure 7B. Likewise, the more difficult items took longer to be answered by the students (Figura 7A). Also, no linear relationship was found between difficulty and the average of correct answers to the items (Figure 7C), a property of adaptive tests that is markedly different from conventional tests, in which there is a high correlation between the average of correct answers to an item and its difficulty estimated via IRT. In general, the three results confirm what was to be expected for a CAT.

FIGURE 7 – Relationship between item difficulty (parameter b) and three aspects of item administration: A) average time to answer the item; B) average proficiency of the students presented with the item; C) average of correct answers to the item. Students in the 1st and 2nd grades of primary education



Source: Prepared by the authors.

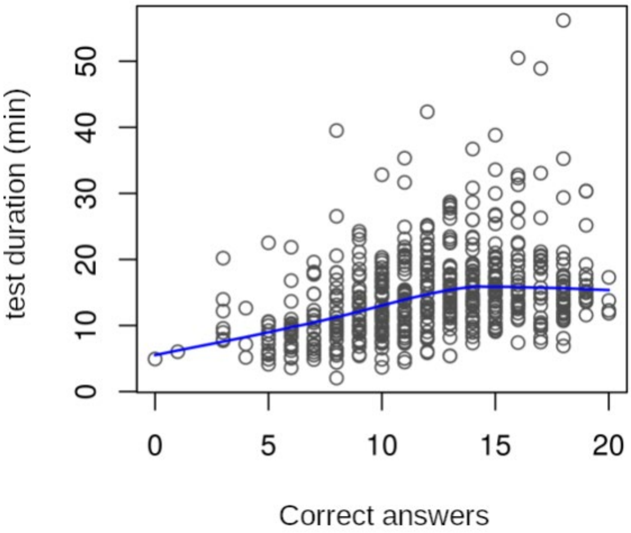
Time Management and Proficiency

Test solving time was also investigated by comparing students of different proficiency levels. The initial hypothesis is that more proficient students tend to manage test time better.

Firstly, the results of the non-adaptive tests[10] show a positive association between proficiency and test solving time (Figure 8). In other words, there is a tendency for more proficient students to spend more time taking the test. Linear regression confirms that this relationship is significant (p <0.001). On the other hand, there seems to be a limit to this tendency, since after 14 correct answers students do not seem to need more time to achieve good results. There is even a slight downward tendency from this point on, tough linear regression indicates it is not significant (p = 0.42).

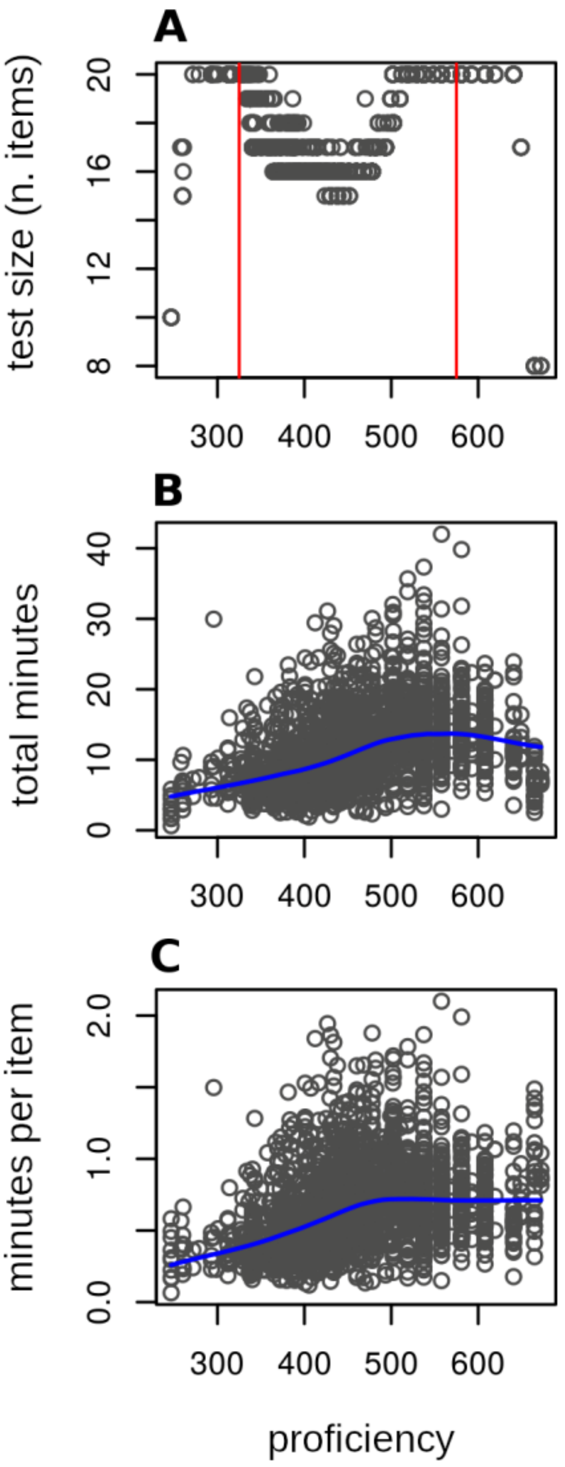
[10] The results and sample of the non-adaptive electronic tests administered in this project as a kind of control are detailed in Alavarse et al., 2018.

FIGURE 8 – Relationship between the number of correct answers and test solving time for Provinha Brasil – Leitura (non-adaptive electronic version). Students in the 2nd grade primary education



Source: Prepared by the authors.

FIGURE 9 – Relationship between proficiency and (A) size; (B) total solving time; and (C) weighted solving time. Students in the 1st and 2nd grades of primary education



Source: Prepared by the authors.
Red lines: minimum and maximum cut-off points. Blue lines: non-linear tendency calculated with estimator M.

The CAT's results are similar. It is worth considering that the CAT has different sizes (in number of items), unlike the non-adaptive test, which always has 20 items. Therefore, analyzing the relationship between proficiency and test solving time is a little more complex with the CAT. Figure 9 summarizes the main information necessary in this regard.

The smallest tests (with eight or ten items) correspond to students at the extremes of the scale shown in Figure 9A. In these cases, the test was terminated by the criterion of *proficiency level reliability*. However, most terminations generated by the adaptive algorithm were based on the *proficiency reliability* criterion, which ended up having an effect on the midmost range in the proficiency scale – where the estimation error is usually smaller. This distribution over the proficiency scale confirms the complementariness of the two test termination criteria in a CAT.

Figure 9B shows a growth tendency up to a certain point. From 600 points onward, there is a slight decline. However, this figure represents the test's total solving time, but the CAT has different sizes. It may be that students with over 600 points spent little time in the test, for the algorithm terminated it quickly. Indeed, Figure 9C confirms this hypothesis, since the decline disappears when we analyze time per item, rather than the test's total time.

Also, the tendency in Figure 9C is similar to that in Figure 8. Indeed, this relationship between proficiency and time dedicated to the test (a positive association up to a certain point, after which it stabilizes) is consistent in both the adaptive and the non-adaptive tests. A significant detail is that test time stops increasing when proficiency is a little higher than the average observed: 463 points in Figure 9C and 13 correct answers in Figure 8. The interpretation of these facts is not clear, but their repetition in both cases indicates fertile ground for research.

A second aspect of the relationship between time management and proficiency was also analyzed. To that end, we selected two sample strata: the lower third and the upper third of students in terms of proficiency. Table 3 compares the two strata regarding some characteristics of the items administered.

TABLE 3 – Mean and coefficient of variation of item solving time according to the strata defined by the lower and upper thirds of students in the 1st and 2nd grades of primary education, in terms of proficiency

STRATA	MEAN SOLVING TIME (MINUTES)	COEFFICIENT OF VARIATION OF SOLVING TIME	TOTAL OF RESPONDED ITEMS
Lower Third	0.45	0.91	29
Upper Third	0.68	0.74	22
P-value (paired t-test)	0.01	0.01	---

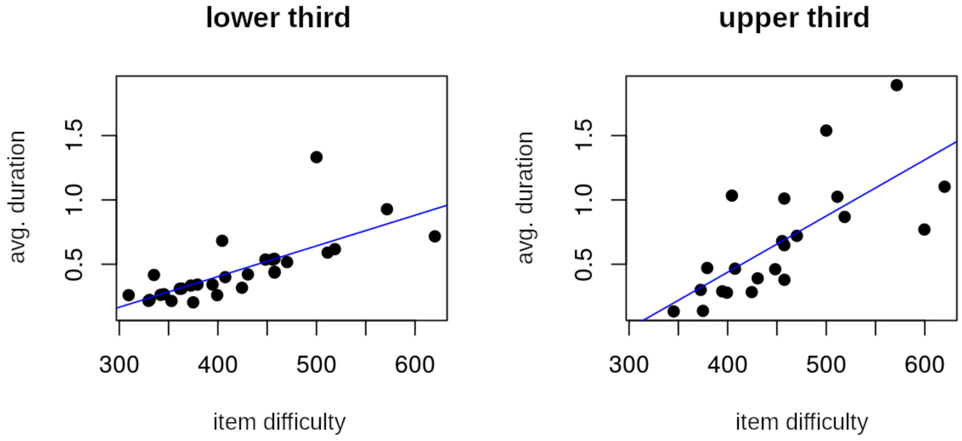
Source: Prepared by the authors.

Note: The significance of the difference between the strata is indicated by the P value of the paired t-test. * Total number of items responded by students in each third.

To determine whether the differences in the means of the two strata described in Tabela 3 are significant, we performed a paired t-test with 95% confidence. The students in the upper stratum tend to spend more time solving the items, which confirms the previous results. In addition, the coefficient of variation indicates that there is less variability in the upper stratum; in other words, these students tend to manage time more homogeneously in each item, thus suggesting a certain rationality regarding their use of time, and perhaps also a more sustained ability to concentrate.

Finally, there is a point to be examined, based on the relationship observed between the average solving time of items and their difficulty (Figure 7A): the students tend to take longer to answer more difficult items, which makes sense and can be considered part of the ability to manage test time. But to what extent is that different in the two strata? Figure 10 shows that, in the upper stratum, the relationship between item solving time and difficulty is stronger than in the lower stratum. Linear regression confirms this.

FIGURE 10 – Relationship between item solving time (in minutes) and difficulty (in Provinha Brasil – Leitura's scale) in each stratum. Students in the 1st and 2nd grades of primary education



Source: Prepared by the authors.
Note: Blue line estimated by linear regression (p<0,001).

In sum, our results suggest that more proficient students tend to manage their time better in the following aspects: a) they dedicate more time to the test in general; b) they dedicate more time to the more difficult items, and less time to easier ones; c) they dedicate time to each item more homogeneously – which could be related, we speculate, to a more sustained ability to concentrate.

CONCLUSION

Our results confirm the quality of the CAT of Provinha Brasil – Leitura and of the methods used to develop the algorithm, starting from the conceptual framework of IRT, and based on simulations and free open source software. Indeed, this text could serve as a model for the construction of other, similar adaptive tests. Several characteristics expected in a CAT were observed, such as a reduction in average test size of and in average solving time, the differential use of items and the absence of a correlation between difficulty and mean of correct answers. The TAC of Provinha Brasil – Leitura also proved to be a useful tool for administration to different student populations, and it may probably also include the 3rd grade, especially by expanding its pool with items that could meet students with greater proficiency, as expected for that grade; though there may be those who perform similarly to students in earlier grades. With regard to 1st graders, because the CAT was administered at the end of the school year, and considering that these students would take Test 1 at the beginning of the next school year, the item pool was suitable. In addition, the CAT's object of assessment – literacy – relates to a competency that is worked on since the 1st grade and is cumulative. Its cumulative nature, less restricted to each grade, was evidenced in the spectrum of average proficiencies observed for the 1st and 2nd grades, thus confirming, in psychometric terms, the quality of Provinha Brasil – Leitura, despite its being developed by INEP mainly for use by teachers of the 2nd grade of primary education. It is also worth highlighting that no 1st grade student expressed "surprise" at the items' contents, considering that all administrations were monitored and documented in reports.

The test termination criterion proposed in this study (proficiency level reliability) also proved effective and potentially useful for tests with a discretized scale (i.e., in levels), which prioritize assessment (not just measurement) and pedagogical interpretation, whether for formative or summative purposes. It is worth noting that the degree of influence of this criterion depends directly on the number of cut-off points and their distribution over the proficiency scale.

In addition, our analysis of test solving time confirms the hypothesis that more proficient students tend to manage test time better. Firstly, there is a positive association between proficiency and test time, which stabilizes in the upper portion of the proficiency scale. That association was observed in both the adaptive and the non-adaptive tests. Another general tendency observed is that the more difficult items take longer to be answered. Comparing the upper and lower thirds of students in the proficiency scale, we observed that: a) the upper third takes longer to answer the items; b) the upper third shows less variation in the time dedicated to each item, which could be related – we speculate – to a certain steadiness in their ability to concentrate; c) the upper third shows a greater slope in the relationship between the item's difficulty and test solving time; in other words, these students tend to spend more time on the more difficult items, and less time on easier ones. This aspect of test time management was found in the entire population, but it is especially present in more proficient students.

It is important to highlight some limitations of this study. Firstly, the simulations did not include the examinee's personal engagement aspect, only examinee proficiency and the item pool's parameters. However, a greater engagement by students is to be expected in an adaptive test, especially by students at the upper and lower extremes of the proficiency scale. Secondly, the estimation method used (WBS) is Bayesian, which yields less accurate results when the population's mean is not close to the expected a priori. In this regard, a non-Bayesian method is recommended, such as weighted likelihood (WL), if there is no reliable information about the population. Another alternative is a mixed method, i.e., starting the test with WL and terminating it with EAP. Thirdly, the item pool used was small and, as seen earlier, the exposure rate for some items was significantly high, while others were not even administered. To overcome this limitation, another item selection method is recommended, since Fisher's Maximum Information tends to generate this type of result, especially when associated with a logistic model with a constant (slope) parameter, as was the case. Another limitation of this study, which is particularly important, is the small size of the item pool; it is advisable to use a pool with at least 100 items for an effective performance by the adaptive test.

Finally, it is worth mentioning possible future improvements, such the inclusion of methods to optimize item pool security (e.g., by controlling the exposure rate), the transformation of the CAT into an MST (i.e., Multistage Test, in which the items are selected in groups), or even the possibility of pre-testing new items during the test's administration.

REFERENCES

ALAVARSE, O.; CATALANI, E.; MENEGHETTI, D.; TRAVITZKI, R. Teste Adaptativo Informatizado como recurso tecnológico para alfabetização inicial. *Revista Iberoamericana de Sistemas, Cibernética e Informática*, v. 15, n. 3, p. 68-78, 2018.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. *Teoria da resposta ao item*: conceitos e aplicações. São Paulo: Associação Brasileira de Estatística, 2000.

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, p. 1-18, Dec. 2012.

BAKER, F. B. *The basics of Item Response Theory*. 2nd ed. Washington: ERIC Clearinghouse on Assessment and Evaluation, 2001.

BARRADA, J. R. A method for the comparison of Item Selection Rules in Computerized Adaptive Testing. *Applied Psychological Measurement*, v. 34, n. 6, p. 438-452, 2010.

BARRADA, J. R. Tests adaptativos informatizados: una perspectiva general. *Anales de Psicología*, v. 28, n. 1, p. 289-302, 2012.

BIRNBAUM, A. Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, v. 6, p. 258-276, 1969.

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.

BRASIL. Ministério da Educação. Portaria Normativa n. 10, de 24 de abril de 2007. *Diário Oficial da União*, Brasília, 26 abr. 2007a. Disponível em: http://portal.mec.gov.br/arquivos/pdf/provinha.pdf. Access on: Aug. 20, 2020.

BRASIL. Presidência da República. Casa Civil. Subchefia para Assuntos Jurídicos. Decreto n. 6.094, de 24 de abril de 2007. Dispõe sobre a implementação do Plano de Metas Compromisso Todos pela Educação, pela União Federal, em regime de colaboração com Municípios, Distrito Federal e Estados, e a participação das famílias e da comunidade, mediante programas e ações de assistência técnica e financeira, visando a mobilização social pela melhoria da qualidade da educação básica. *Diário Oficial da União*, Brasília, 25 abr. 2007b. p. 5. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2007/Decreto/D6094.htm. Access on: Aug. 20, 2020.

BRASIL. Ministério da Educação. *Provinha Brasil*: avaliando a alfabetização: guia de apresentação, correção e interpretação dos resultados: Teste 2. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2016.

CATALANI, Érica Maria Toledo. *Teste Adaptativo Informatizado da Provinha Brasil*: a construção de um instrumento de apoio para professores(as) e gestores(as) de escolas. 201. 282 f. Tese (Doutorado em Educação) – Faculdade de Educação, Universidade de São Paulo, São Paulo, 2019.

FERREIRA, D. F. *Estatística básica*. Lavras, MG: Ed. UFLA, 2005.

GEORGIADOU, E. G.; TRIANTAFILLOU, E.; ECONOMIDES, A. A. A Review of Item Exposure Control Strategies for Computerized Adaptive Testing developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment*, v. 5, n. 8, p. 1-38, 2007.

GONTIJO, C. M. M. Avaliação da alfabetização: Provinha Brasil. *Educação e Pesquisa*, v. 38, n. 3, p. 603-622, 2012.

LINDEN, W. J.; GLAS, C. A. W. *Elements of Adaptive Testing*. New York: Springer, 2010.

LORD, F. *Applications of Item Response Theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates, 1980.

MAGIS, D.; RAÎCHE, G. Random generation of response patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software*, v. 48, n. 8, p. 1-31, 2012.

MORAIS, A. G.; LEAL, T. F.; ALBUQUERQUE, E. B. C. "Provinha Brasil": monitoramento da aprendizagem e formulação de políticas educacionais. *Revista Brasileira de Política e Administração da Educação*, v. 25, n. 3, p. 301-320, maio/ago. 2009.

PARTCHEV, I. *irtoys*: A Collection of Functions Related to Item Response Theory (IRT). S.l.: The Comprehensive R Archive Network, 2016. Disponível em: https://cran.r-project.org/package=irtoys. Access on: Mar. 19, 2020.

REIF, M. *PP*: Estimation of person parameters for the 1,2,3,4-PL model and the GPCM. S.l: GitHub, 6 abr. 2019. Disponível em: https://github.com/manuelreif/PP. Access on: Mar. 19, 2020.

REVUELTA J.; PONSODA, V. A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, v. 35, n. 4, p. 311-327, 1998.

SOARES, M. *Alfabetização*: a questão dos métodos. São Paulo: Contexto, 2016.

WARM, T. A. Weighted likelihood estimation of ability in item response theory. Psychometrika, v. 54, n. 3, p. 427-450, Sept. 1989.

WEISS, D. J.; KINGSBURY, G. G. Application of Computerized Adaptive Testing to educational problems. *Journal of Educational Measurement*, v. 21, n. 4, p. 361-375, Winter 1984.

NOTE: This article was collaboratively written by the authors. Rodrigo Travitzki created the algorithm, Ocimar Munhoz Alavarse was the project's general coordinator, Douglas De Rizzo Meneghetti created the electronic test's administration software and Érica Maria de Toledo Catalani conducted the psychometric analysis of results.

Received on: March 19, 2020
Approved for publication on: September 16, 2020