Informatics in Education, 2018 Vol. 17, No. 2, 381–404

DOI: 10.15388/infedu.2018.20

An Adaptive Test Analysis Based on Students' Motivation

Sérgio R. I. YOSHIOKA, Lucila ISHITANI*

Graduate Program in Informatics, Pontifical Catholic University of Minas Gerais (PUC Minas)
e-mail: sergio.yoshioka@sga.pucminas.br, lucila@pucminas.br

Received: March 2018

Abstract. Computerized Adaptive Testing (CAT) is now widely used. However, inserting new items into the question bank of a CAT requires a great effort that makes impractical the wide application of CAT in classroom teaching. One solution would be to use the tacit knowledge of the teachers or experts for a pre-classification and calibrate during the execution of tests with these items. Thus, this research consists of a comparative case study between a Stratified Adaptive Test (SAT), based on the tacit knowledge of a teacher, and a CAT based on Item Response Theory (IRT). The tests were applied in seven Computer Networks courses. The results indicate that levels of *anxiety* expressed in the use of the SAT were better than those using the CAT, in addition to being simpler to implement. In this way, it is recommended the implementation of a SAT, where the strata are initially based on the tacit knowledge of the teacher and later, as a result of an IRT calibration.

Keywords: computerized adaptive test, stratified adaptive test, item response theory, motivation, anxiety, evaluation methodologies, teaching/learning strategies, improving classroom teaching.

1. Introduction

Advances in digital Information and Communication Technologies (ICT) have significantly changed several areas of knowledge, including Education. In the current context, students have access to various resources and information available on the web, unlike previous generations when the teacher was the central person and held most of the knowledge. To make better use of this potential, teachers are increasingly required to play the role of mentor, using their experience to recommend good content to their students, leaving them free to explore the space of knowledge. Thus, to support the teacher, it is interesting to develop intelligent tools, which adapt individually to each student, identifying their strengths and weaknesses and reporting their results.

* Corresponding author

In recent years, adaptive software has become widely used, especially in the United States (Pardos *et al.*, 2012). Among these, Computerized Adaptive Testing (CAT) is a computerized test that adapts to the user's knowledge, avoiding boredom or frustration, and evaluating more quickly and accurately than traditional tests (Meijer and Nering, 1999), whose set of applied questions are fixed and equal for all. An interesting and well-founded way of implementing CAT is to apply the Item Response Theory (IRT) as it enables students who have taken different tests to be compared by the same scale.

IRT emerged in the 1930s, but gained more power with the work of Lord (1952), which later updated by Birnbaum (1968). One of the most used models relates the ability of a respondent to his or her chance of being able to answer a question by three parameters: discrimination, difficulty, and pseudo-guessing. The first parameter refers to how well a question discriminates apt students from unfit ones. The second is related to the difficulty of the question itself. And finally, the last is related to the chance of guessing the correct answer (Section 2.1).

Pilot parameters and statistical techniques are used to infer the parameters of the items (a process known as calibration) that must be applied to all issues that make up the bank of CAT items, often requiring specialized teams in statistics and computer science. Since it is necessary periodically to elaborate and insert new items to ensure the reliability of the test, this task, which was already costly, becomes even more so.

To minimize this effort, some dynamic calibration approaches are presented in the literature, such as the CBAT-2 algorithm (Huang, 1996) used in CALEAP-Web (Piton-Gonçalves *et al.*, 2009). However, in the general context of face-to-face teaching, classes contain only a few dozen students, which leads to low sampling and possibly incorrect calibration, affecting CAT accuracy.

In contrast to the IRT-based CAT, which in this work will be called CAT only, there is the stratified adaptive methodology (González-Sacristán *et al.*, 2013), which in this work will be called SAT. Although not based on the IRT, SAT aims to improve the quality of the assessment, to have greater robustness, safety and relevance of the presented questions, besides being easy to implement and to apply in the daily context of a classroom. Its advantages, compared to CAT, consist of less effort, dispensing with the team of statisticians, besides the ease of implementation. However, the concept of the difficulty of a question is subjective and depends on the tacit knowledge of the specialist who will classify the strata of questions (Section 2.3).

Using adaptive tests, it is expected to avoid boredom in students of high ability or frustration in those of low ability. This phenomenon has already been explored in other contexts, such as in Flow theory (Csikszentmihalyi, 2000) and also in the Inverted U Theory (Yerkes and Dodson, 1908). However, the motivation of the student in the use of adaptive tests is little explored in the literature. In addition, this aspect is even more relevant if the goal is to support learning, not only to evaluate (Section 2.4), since learning takes place over longer period than assessment and motivation can encourage or discourage students to continue studying. In addition, there is evidence in the literature that motivation significantly influences student achievement (Section 2.5).

Thus, this work presents a case study of a question selection method (González-Sacristán *et al*., 2013), compared to CAT, using seven classes of the Computer Networks course, in 11 experiments. For this, we used ATES, a system developed for this research, which collected data on student's motivation, through the Questionnaire on Current Motivation (Vollmeyer and Rheinberg, 2006). In addition, we also empirically evaluated the best time to abandon a specialist calibration in order to use the estimated parameters of a new item inserted in an item bank considering a CAT dynamic calibration.

The results indicate that the levels of *anxiety* declared by those who used SAT were better than those who used CAT. Besides, SAT is simpler to be implemented and therefore, among these two, the most recommended would be the use of SAT, in the context applied to face-to-face teaching. One hypothesis that arises to explain higher anxiety is that CAT can converge to very difficult issues at an early stage of the activity, negatively impressing the student. However, it has been found that adaptability can positively influence anxiety levels, since students tend to declare themselves increasingly anxious as the difference between the difficulty of a question and their ability becomes greater. In short: being adaptive is positive, converging quickly to very difficult issues isn't.

This paper is organized as follows: Section 2 presents the background of this work, which includes the main characteristics of IRT, CAT and SAT, and also related work about students' motivation while carrying out tests. Section 3 presents the adopted methods. Section 4 presents the main results, and Section 5 our main conclusions and suggestions for future work.

2. Background

This section presents the main characteristics of IRT, CAT and SAT. In addition, it presents types of tests and discusses related work on students' motivation while taking a test.

2.1. *Item Response Theory (IRT)*

According to Hambleton *et al*. (1991), IRT consists of a family of mathematical models that explain a person's performance on a test item through their latent traits, or abilities. Hambleton *et al.* (1991) also stated the IRT was proposed in order to solve some limitations that occur in the classical methods of evaluation. In the classical model, the characteristics of the assessed group are mixed with the characteristics of the evaluation: they can not be separated. If the evaluation contains many issues of low difficulty, the assessed group has better results and may be considered to have high ability; if it is very difficult, the group has worse results and may be considered to have low ability. Similarly, for the same evaluation, if we give it to a group of students with low ability, it will be perceived as difficult; if we give it to a high skill group, the evaluation will be perceived as easy. In this way, the task of comparing two populations that performed two different tests, even on the same subject, becomes complicated.

IRT is based on two assumptions (Hambleton *et al.*, 1991, Couto and Primi, 2011). The first one is unidimensionality, at which the test is presumed to measure only one ability (or dimension). There is a consensus that several skills are used in the execution of a task, however it is understood that there is a dominant ability among the several ones that influence behavior. So, this dominant ability is the one that will be measured in the IRT. Assuming that more than one ability is required to respond to an item, multidimensional models are proposed. However, these models present high implementation complexity.

The second assumption, the local independence, assumes that the response to an item will not influence or cause a variation in the ability of the student. In other words, the probability of guessing a correct response is independent of the response to a previous item. In this way, attention and care are necessary so that the item bank maintains a high cohesion, to avoid violation of the one-dimensionality and local independence.

2.2. *Computerized Adaptive Testing (CAT)*

Oppermann *et al.* (1997) distinguish *adaptivity* from *adaptability*. The first refers to systems with characteristics that adapt to users by themselves, while the second refers to systems that require users to inform their preferences. So, Computerized Adaptive Testing (CAT) refers to tests that have self-adaptive characteristics and are supported by information technology.

CAT aims to tailor a test to the student according to her ability. For each question answered, the system selects from the item bank the next appropriate question according to the answers given previously. It is expected that very simple items can be boring and very complex items can frustrate, negatively influencing learning (Section 2.5). An interesting way to select appropriate questions is through the IRT model, which proposes to describe the behavior of a person when responding to an item. Besides maximizing the benefits of CAT, IRT allows different people to receive different tests with different sizes (Meijer and Nering, 1999).

Meijer and Nering (1999) list some of the advantages and disadvantages of CAT over the classic pen and paper test. As advantages, tests can be shorter, with better accuracy, customized, with immediate results and also with the possibility of automatic generation of reports. It usually requires fewer items to evaluate, because conventional tests need to contain issues of varying degrees of difficulty, while the CAT directs and focuses on the appropriate difficulty of the test, resulting in shorter test times (Weiss and Bock, 1983, Lord, 1980).

Among the disadvantages, it has a high cost of implementation, requires a specialized team to organize the software, besides needing an updated item bank in order to keep the test valid.

According to Thompson and Weiss (2011), the components of CAT are:

- An item bank or set of questions, usually already calibrated, that is, with its known parameters.
- A criterion for choosing the first item.
- An item selection algorithm: given the user's estimated ability, the algorithm will choose the next appropriate question.
- An ability estimation algorithm: given the user's response, the algorithm will update its estimated ability.
- A stop criterion.

Among these components, the item bank is the most important to the success of the CAT, since it requires careful preparation of questions and initial calibration (Manseira and Misaghi, 2013). As previously described in Section 2.1, it is necessary to elaborate high cohesion items, which revolve around the same competence, in order to maintain the presumption of unidimensionality and also the absence of items that provide tips for others, so as to keep local independence. These precautions are necessary for a good IRT estimation.

In addition to these precautions, there are some problems regarding item calibration, or item parameter discovery. Frequently the parameters are discovered by the pilot tests for a sample of the population in which the test is desired to be applied. After that, one statistical method is used, among which the most used are Maximum Likelihood and Bayesian methods such as Markov Chain Monte Carlo (Ricarte, 2013). All of this has a high cost for organizations. And the need for specialized teams in statistics and computer science can make CAT unfeasible for day-to-day use. There is also the need to constantly update the item bank, since items may leak between classes making the test invalid. This makes the CAT deployment cost permanent throughout its lifespan.

To minimize these disadvantages in CAT deployment, dynamic calibration approaches are proposed, such as dispensing with the initial calibration. In this context, Huang (1996) presented the CBAT-2 algorithm: the questions are previously ordered by experts and the parameters are constantly checked as the questions are answered. Hirose and Aizawa (2014) proposed the use of two databases: one to select the item and the other to calibrate them. These banks are daily synchronized. In addition, synthetic data insertion methods are used to increase the volume of data and speed up the calibration of items.

2.3. *Stratified Adaptive Test (SAT)*

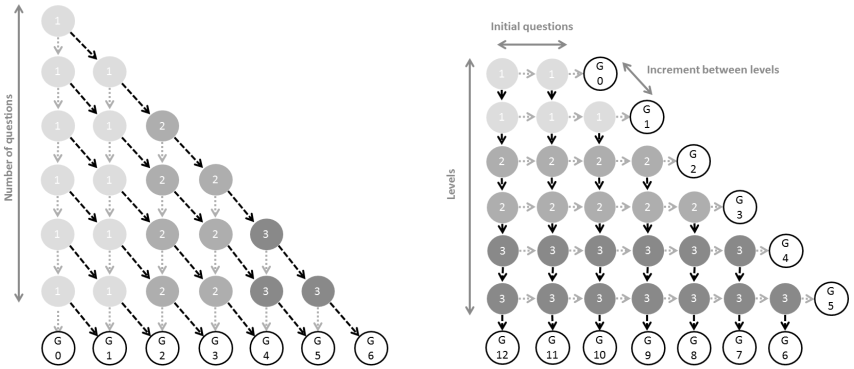
Besides CAT, other adaptive models are proposed in the literature, such as an adaptive test based on rating or hit rate (Silva and Direne, 2016) or the proposal of Tesseroli *et al.* (2016), which determined the complexity of an item through the concepts it is related with, organized the items in a graph and searched for a heuristic to select the increasing complexity of the items. Another option is the Stratified Adaptive Test (SAT), a term that was first used by Betz and Weiss (1976) and reappeared as a new approach in (González-Sacristán *et al.*, 2013).

González-Sacristán *et al*. (2013) proposed a computationally assisted method to improve the quality of the evaluations, increasing the robustness, objectivity, safety and relevance of the content. The approach, inspired by CAT, organizes the items into categories ordered by the most basic and fundamental to the most specific and elaborate, which is only relevant after mastering the knowledge of the previous categories. With this model, González-Sacristán *et al*. (2013) want to better reward students who invest on a solid foundation than those who prefer to go directly to the advanced levels without having the robust mastery of fundamental concepts. They presented two models. In Model 1 (Fig. 1a), each student answers the same number of questions. One right answer change to the next question diagonally and an incorrect answer, vertically. In this way, it is only possible to answer questions from higher levels after correctly answering some basic questions. This walking continues until a leaf node G_n is reached, which indicates the result obtained by the student.

In Model 2 (Fig. 1b), the amount of answered questions varies. In this model, an incorrect answer moves horizontally, while a correct one moves vertically. A sequence of erroneous responses may lead to an early ending of the test. Like Model 1, the leaves G_n represent the end of the test and the result obtained by the student.

So, González-Sacristán *et al*. (2013) assessed whether the two models discriminated well four kinds of students, performing numerical simulations. The possibilities are: good student (student who goes well on all levels), bad (student who goes badly on all levels), direct (student who gives more importance to fundamentals), inverse (student who focuses on advanced questions). The tests were performed with two types of questions: multiple choice and completion. The results indicated that the completion test is more adequate to identify the student profiles.

Finally, to evaluate if the model is adequate in the evaluation of basic programming skills, the authors carried out a study with approximately 100 undergraduate



(a) Model 1

(b) Model 2

Fig. 1. Item selection models.
Source: (González-Sacristán *et al*., 2013)

students in 2012 and 2013. They applied two tests using the proposed methodology: a completion test and an open answer test. After the students who did not perform both tests were removed from the study, a Pearson coefficient of 0.66 and Spearman of 0.62 was found between the results of both tests, indicating that the completion questions, which are possible to be computationally assisted, are a good substitute for open questions.

2.4. *Tests and Learning Improvement*

Behar (2009) and Jorba and Sanmarti (2003) present three types of assesments: diagnostic, formative, and summative. The first refers to the one applied at the beginning of the learning process and aims to obtain information about the previous knowledge of the student. The second one refers to the one performed throughout the process, assuming that the students can learn by restructuring the knowledge from the activities that they perform (Jorba and Sanmarti, 2003). And the last occurs at the end of the process, in order to verify if the proposed objectives were reached or not by the students.

Until recently, it was mistakenly believed that the tests provided a formative evaluation in order to measure knowledge, but they contributed little to learning, which would only occur while studying (Yeh and Park, 2015). Contrary to what was believed, some recent research has shown that tests do significantly help with retention of knowledge, and this encourages learning.

This principle is called *Test Effect* (Yeh and Park, 2015). In another interesting phenomenon, the *Spacing Effect* (Ebbinghaus, 1885, Yeh and Park, 2015), the best retention of long-term knowledge is better for those who opt for pauses between studies, rather than massive repetitions. This is interrelated to the *Test Effect* and can be automated by a computerized test.

Karpicke and Roediger (2008) observed that studying repeatedly after a formal test did not bring any benefit, but performing tests periodically slowed down the inevitable forgetfulness. They also noted that feedback with a result other than expected (failure in an item recognized as easy or success in an item recognized as difficult) increases the likelihood of reinforcement and retention of knowledge.

Still on this topic, Finley *et al.* (2011) stated that a very long space of time between studies makes the task of remembering very arduous, so the chances of forgetting are great. For an item to be retained with optimum power, ideally it should be retrieved shortly before it would be forgotten. Landauer and Bjork (1978) demonstrate that scheduling reviews with short intervals at the beginning of the learning and gradually spacing more, favors a better compatibility with the optimal retention than reviews with uniform intervals. This is due to a principle similar to CAT: effort is lower in the initial stages of learning and difficult as knowledge is mastered. By this principle, another approach to manipulating the desired difficulty is to vary the amount of course material available to the student to respond to an item.

2.5. *Tests and Student Motivations*

Motivation is “the internal process that energizes, directs, and sustains behavior” (Reeve, 2014). MacIntyre (2002) states that almost all human behavior is motivated by something. However, motivation can suffer from positive external stimuli (motivation) or negative (demotivation) stimuli. In this sense, this section brings together works related to favoring student motivation when using computerized tests.

Lilley et al. (2005) sought to quantify the difficulty perceived by students after the application of a CAT. Twenty-four items on Human Machine Interaction (IHC) were applied. At the end of the evaluation session, participants indicated the difficulty of the test, using the scale: 1 (very easy) to 5 (very difficult). After that, through a Kruskal-Wallis statistical test, no correlation was identified between student performance and perceived level of difficulty, providing evidence that CAT is effective in adapting the items presented for each level of individual proficiency.

Karadeniz (2011) studied the relationships between the students’ gender and their anxiety in performing a mobile-mediated assessment. Ten questions about computer hardware were applied to 20 second-year high school students. Anxiety levels were measured with the *Motivated Strategies for Learning Questionnaire* (MSLQ) (Pintrich *et al*., 1993). The results indicated that the group of students with lower level of anxiety performed better than the more anxious group.

Jansen *et al*. (2013) studied the practice of mathematics with CAT and its benefits for anxiety, perceived competence and performance. 207 children were randomly organized into four groups, according to their pre-test performance.The groups were: “hard” (60% hit), “medium” (75% hit) and “easy” (90% hit), besides the control group that did not used CAT. Excluding the latter, the Math Garden system was used 3–5 times per week with each session lasting approximately 10–15 minutes during 6 weeks. Pre and post tests were based on the use of *Math Anxiety Scale for Children* (MASC) (Chiu and Henry, 1990), *Perceived Competence Scale for Children* (Veerman *et al*., 1997) and *TempoTest Automatiseren* (TTA) (De Vos, 2010). The results indicated that, in the end, all four groups had lower levels of anxiety compared to the pre-test, and there were no significant differences between the groups. The “easy” group achieved significant improvement over the control group, while the “medium” and “difficult” groups showed improvement, but not significant. In general, Jansen *et al*. (2013) indicate the use of CAT for teaching mathematics, especially those who implement features of increasing success rate and privacy when students make mistakes.

Penk and Schipolowski (2015) investigated expectation, value and effort, their interrelationships, and their relationship to performance in a large-scale evaluation. The expectation refers to how students notice their own performance; value refers to the perceived benefit coming from the test, such as importance, pleasure, and utility; and effort is the component that refers to the cost, or how much effort is necessary to be successful in the test. Motivation measures were collected before and after the tests by the Questionnaire on Current Motivation (QCM) (Freund and Holling, 2011). Three items of the Test Motivation Scale (Eklöf, 2010) were used to measure the *Effort*. The most relevant factors were *Effort* and *Expected Success*.

Ortner *et al.* (2014) reviewed the literature about student motivation when using CAT in relation to traditional tests. Among all the cited paper, (ggen and Verschoor (2006) stated that CAT may be perceived as more difficult as students are used to traditional tests with a higher likelihood of success.

Ortner *et al.* (2013) also observed an unexpected fact: in one CAT, the students felt more satisfied after conducting tests of less difficulty, although they obtained worse results than in more difficult tests. Based on this, Ortner *et al.* (2014) conducted an experiment to investigate the motivational effect of CAT. For a group of 174 high school students, divided into two groups, a matrix test (Hornkes *et al.*, 2000) was applied in both modes: adaptive and traditional. Five minutes after the initial time of the test, it was interrupted in order to measure the Flow (Csikszentmihalyi, 2000), the fear of failure and the expected success. For collecting these data, it was used the questionnaires of Vollmeyer and Rheinberg (2006). The results showed that there was a greater fear of failure and less expected success in CAT. However, regarding Flow, there were no significant differences when comparing the two environments. It was also observed that the greatest measures of Flow were related to the students with better performance.

For this research, a literature review on the motivation mediated by computerized tests was carried out. Among 1077 articles searched in three digital libraries (Science Direct, ACM Digital Library and IEEE Xplore Digital Library), only six papers addressed this topic, of which only two were related to CAT. This demonstrates the low amount of research in this area versus the amount of recent CAT-related work. In addition, the selected papers pointed out that items of low difficulty have more significant improvements in performance (Jansen *et al.*, 2013, Ortner *et al.*, 2014), and that motivation is a relevant performance factor, with evidence of influencing up to 28% in the variance of performance (Penk and Schipolowski, 2015). Also, there is a correlation between motivation and performance (Karadeniz, 2011, Penk and Schipolowski, 2015, Ortner *et al.*, 2014), justifying the importance of further research in this area. Another aspect to note is that these papers investigate large-scale contexts. This scenario diverges from the usual classroom, where there may be only a few dozen students, small statistical sampling and a shortage of teams specialized in statistics and computer science.

3. Method

This section presents the activities conducted to obtain the results presented in this paper.

3.1. *Composition of the Item Bank*

Considering the topics of a course on Networks (Table 1), a professor of this course selected the items to be applied by collecting them from public tenders and also elaborating some questions by himself. These questions had multiple choice format and only one correct choice. They were ordered in degrees of difficulty: *very easy, easy, regular, difficult, very difficult*. *Very easy* difficulty corresponds to the questions with more fun-

Table 1

Topics in item bank

Code	Topic
A1	Introduction
A2	Physical layer
A3	Data link layer
A4	Transport layer
A5	Application layer
A6	Network layer

Table 2

Number of questions per difficulty level

Topics	Very easy	Easy	Regular	Difficult	Very difficult
A1 e A2	5	4	3	2	1
A3	5	4	4	2	1
A4 e A5	5	4	3	2	1
A6	5	4	3	2	1

damental and elementary concepts of the course. The number of questions elaborated and organized by difficulty, as recommended by the professor of the course, is described in Table 2.

3.2. *Selection of the Methods to be Compared*

This research considered and compared two methodologies of adaptive selection of items: SAT (González-Sacristán *et al*., 2013) and IRT-based CAT (Section 2.2).

Considering SAT, we chose Model 1 (Section 2.3), because our intention is to use tests for study and preparation and not just for evaluation. In this way, Model 2 present a smaller number of questions to be executed when the student has low ability, taking from her the opportunity to exercise and to foment her curiosity on the topics of a course. Considering CAT, the items were calibrated using PARAM (Section 3.3).

3.3. *Selection of Tools*

We used two tools: ATES and PARAM.

ATES was developed for this study. It is a tool for Web, written in Groovy/Grails. It allows students themselves to register with login and password, to accept the consent term, to execute the questions of the course, to receive feedback and to fill out the Questionnaire on Current Motivation.

PARAM (Rudner, 2012) is an item calibration system for IRT. PARAM implements Newton-Raphson's Maximum Likelihood method as proposed by Lord (1980) and its effectiveness was compared with the BILOG system, presenting similar results (Rudner, 2012).

Given the students' answers, ATES export the response files to the format read by PARAM, which then processes this file and generates an output file. The output file containing the students' ability (θ) can be imported into ATES.

Since in the IRT-based CAT selection method, it is necessary to have the items calibrated beforehand, we opted for initially carrying out the questions in the SAT. This was important to have information to calibrate the questions to be used later.

3.4. *Selection of Test Groups*

As test groups we chose nine classes of Computer Networks courses: seven from the Information Systems program and two from the Computer Engineering program. Table 4 lists all the nine classes. T1-2 is the one of the academic semester subsequent to that of the T1-1 and T2-2 is the one that follows T2-1. More details about the expected competences of courses on Network in Computer Engineering and Information System undergraduate programs can be seen in Table 3 as recommended by Brazilian Comput-

Table 3

Expected competence of course on Network (Araujo *et al*., 2017)

Undergraduate Program	Ref. Code	Competence
Computer engineering	C.2.1	Determine performance and reliability requirements, design, implementation and testing of electronic components and hardware systems.
	C.2.4	To carry out the design of integrated hardware and software for various areas of electro-electronic industry.
	C.3.3	Identify standards, documentation necessary techniques in projects, services and Computer Engineering experiments
	C.3.4	Apply project management methodologies, services and engineering experiments in the area of computing
Information Systems	C.1.1	Decompose the functioning of organizations, social and business systems as Information Systems, distinguishing its elements and multiple internal relations and external models and constructing models for their representation.
	C.3.1	Assess the need to computerize systems, articulating individual and organizational visions, and appreciating opportunities for improvements and/or changes in processes, with the use or evolution of the software.
	C.5.2	Evaluate the physical and logical architecture of communication and computers for organization, using concepts from the reference models, analyzing the operation and performance of their components, applying the concepts of high availability and load balancing, and using virtual machines and management software.

Table 4
Classes

Code	Course	Campus	Professor
T1-1	Information Systems	C1	P1
T2-1	Computer Engineering	C1	P1
T1-2	Information Systems	C1	P1
T2-2	Computer Engineering	C1	P1
T3	Information Systems	C2	P2
T4	Information Systems	C3	P3
T5	Information Systems	C1	P4
T6	Information Systems	C4	P4
T7	Information Systems	C5	P5

Table 5
Data collection

Code	Topics	Classes	Method	B/A
E1	A1 e A2	T1-1 e T2-1	SAT	B
E2	A3	T1-1 e T2-1	SAT	B
E3	A4 e A5	T1-1 e T2-1	SAT	B
E4	A1 e A5	T1-2 e T2-2	CAT	B
E5	A1 e A2	T3	CAT	A
E6	A1 e A5	T4	SAT	B
E7	A1	T5	CAT	-
E8	A3	T6	SAT	-
E9	A1 e A5	T7	SAT	A
E10	A4	T1-2	SAT	A
E11	A6	T1-2	SAT	B
E12	A4 e A6	T2-2	SAT	B
E13	A2 e A3	T1-2 e T2-2	CAT	B

ing Society. Each professor has the autonomy to develop the education plan with its activities and assessments.

In order to select the method and subject of the test of each class, the class should have already been presented to the proposed topics. It is worth emphasizing that although the course is the same, the order in which the topics were presented could be different for each class. A balance was also sought in the number of samples to make comparisons at the time of analysis.

The experiments followed the distribution listed in Table 5, where the topics are those listed in Table 1, the classes are those listed in Table 4, the methods can be IRT-based CAT or SAT and the last column indicates whether the test was performed before (B) or after (A).

3.5. *Data Collection*

Data collection was conducted through the following activities:

- Registration – in ATES, there are two ways to register students: the administrator manually registers the students and allocates them in each class or the administrator activates the available classes and students register by themselves.
- Acceptance of the informed consent form – upon first accessing the system, students must accept an Informed Consent Term already approved by the Research Ethics Committee.
- Solving items – students solve five items per subject following the chosen methodology of adaptive selection of items.
- Filling in QCM – after solving the items, students were asked to fill in the Questionnaire on Current Motivation (QCM) (Vollmeyer and Rheinberg, 2006) adapted for the Network course. The objective was to measure motivation factors for *Anxiety* (A), *Interest* (I), *expected Success* (S) and *Challenge* (C). The available options followed the Likert scale of five options, such that 0 means "strongly disagree" and 4, "strongly agree".
- Feedback for students – after responding to the QCM, the students had access to their results and to the answer key and they were also able to answer the other questions of the bank that were presented to them in increasing order of difficulty. This format was chosen because learning is favored when feedback is not immediate and also when the difficulty is adjusted, starting from the easiest activity to the most difficult ((Yeh and Park, 2015, Finley *et al.*, 2011). The results were summarized by subject showing the number of correct answers.

3.6. *Data Analysis*

Data analysis included a comparative analysis of the motivation factors between the methodologies studied (SAT and IRT-based CAT) and an analysis of the evolution of the difficulty parameter according to the sampling and some calibration strategies.

We used Pearson correlation (Mukaka, 2012) for the analysis of the relations with *Motivation*.

In order to analyze the motivation factors, some scenarios were identified, such as the comparison of two subsequent classes of the same teacher, or groups that took a test on the same subject, but with different methodologies for item selection. For each scenario:

- Similar experiments of this scenario were grouped.
- The mean and the confidence interval of each group of the analyzed scenario were calculated.
- The Pearson correlation between the measured motivation factors, the number of correctly answered questions, and the estimated ability for IRT using the PARAM was calculated.
- The analysis of the scenario was carried out and the recommendation of the adaptive methodology for item selection considering the scenario was produced. In addition, behavioral cues and premises to be tested in future work were also produced.

Considering the collection of recommendations extracted from each scenario and its particularities, we synthesized a recommendation on an adaptive test based on student motivation.

For the evaluation of the difficulty parameter, for each subject, several response files were generated by means of the ATES in such a way that: the first file contains the first ten samples, the subsequent contains the samples from the previous one, plus one; and thus the generation of files follows until reaching the maximum number of samples collected for that subject. In this study, a sample is the set of answers of a student to a certain subject, since she has answered at least five questions of that subject. The samples are organized in chronological order of the execution of the set of questions: from the first one that started responding to the last one that started responding.

After that, each file is processed in PARAM (Section 3.3). The output files (one for each response file) which contain the calibration parameters of each item and also the estimated ability of each respondent were imported into ATES. So, the graphs used in the analysis, are:

- Evolution of the difficulty of the questions by number of samples – it is a line chart, where each line represents the difficulty of an item. The x axis represents the number of samples used in the calibration and the y axis represents the difficulty of the question.
- Mean difference between sample units. – It is a line chart where each line represents the mean difference of the estimated difficulty values for the items between two quantities of consecutive sample.
- Number of crosses between questions per sample unit – It is a bar chart that represents the total number of times that lines intercept two consecutive sample quantities. Crossing lines represents an inversion in the relationship “this is a more difficult question than the other” and therefore has an effect on the decision of which question will be selected adaptively.

4. Results

This section presents the main results. As previously discussed, student participation was voluntary. The number of participants is described in Table 6. It is important to note that not every student who answered the minimum of five questions also answered the Questionnaire on Current Motivation (QCM).

4.1. *Analysis of Motivation Factors*

Scenarios were constructed in order to compare SAT with IRT-based CAT. The graphs presented in this section represent the mean and confidence interval of the skill (θ), *anxiety* (A), *expected success* (S), *interest* (I) and *challenge* (C) of the samples. The measures of *anxiety* (A), *expected success* (S), *interest* (I) and *challenge* (C) were extracted from students’ responses to the Questionnaire on Current Motivation.

Table 6

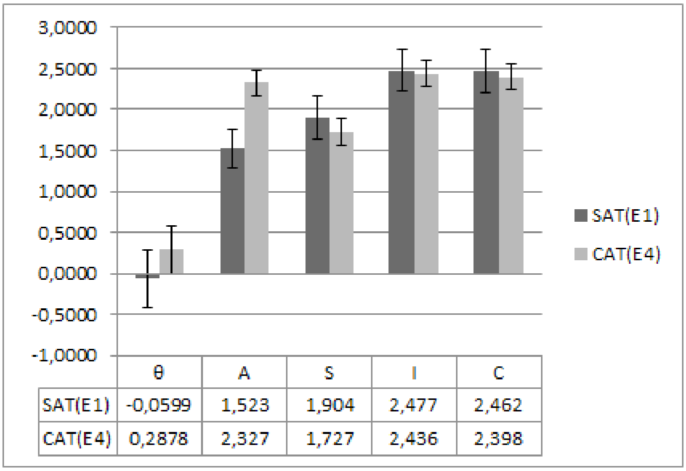
Number of participants

Code	> 5 items	QCM
E1	18	14
E2	11	9
E3	8	7
E4	23	22
E5	12	9
E6	8	7
E7	-	-
E8	-	-
E9	9	9
E10	1	1
E11	16	14
E12	4	4
E13	15	15

4.1.1. *Scenario 1: E1 (SAT) X E4 (CAT)*

In the first scenario (Fig. 2), samples of the experiment E1 and the experiment E4, which used SAT and CAT respectively, were compared. These samples consist of the first module of two subsequent classes of the same professor. However, since there was a change in the order in which the material was presented during the semester, the groups did not answer questions on the same topics.

In this scenario, we observed the proximity of the estimated ability of both groups (difference of 0.33), and that less *anxiety* was registered in the sample that performed the SAT in a confidence interval whose level of confidence is 75%. The measures of *expected success*, *interest* and *challenge* are close and statistically equivalent.



	θ	A	S	I	C
SAT(E1)	-0,0599	1,523	1,904	2,477	2,462
CAT(E4)	0,2878	2,327	1,727	2,436	2,398

Fig. 2. Motivation factors of the Scenario 1.

4.1.2. *Scenario 2: E6+E9 (SAT) X E4 (CAT)*

In the second scenario (Fig. 3), the samples of experiments E6 and E9 using the SAT were compared with those of the experiment E4, which used CAT. These have in common the same subject and, as so, the question bank used for both was the same.

In this scenario, it was observed that the SAT group had the worst performance and presented greater *anxiety*, although expressing greater *interest* and *challenge*, with confidence level of 75%. Another interesting factor is that, despite having a significantly lower performance, the group expressed a greater expectation of *success*, a phenomenon already registered by Ortner *et al*. (2013). These results provide indications, although not generalizable, of a possible relationship of performance inversely proportional to anxiety.

4.1.3. *Scenario 3: E1 (SAT) X E5 (CAT)*

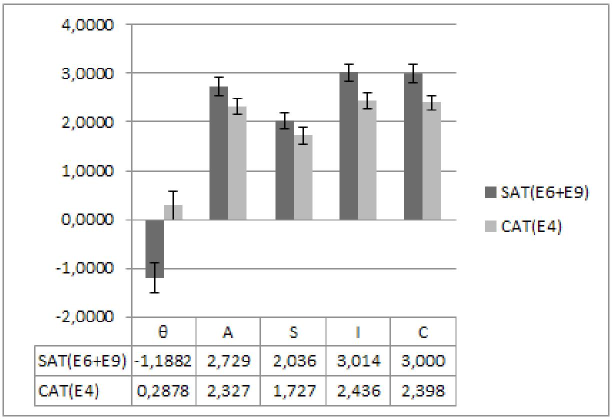
In the third scenario (Fig. 4), we compared the results of samples E1 and E5, which had the same subjects and used the SAT and CAT respectively. In this scenario, CAT samples had poorer ability but lower *anxiety*, contradicting the behavior of the second scenario. However, considering the confidence intervals, the samples had equivalent ability and *anxiety*. A justification for this behavior is that E5 was collected after the formal tests of the course and therefore the students presented lower level of commitment and *interest*, and to that extent, less anxiety. These results provide indications, although not general, of a possible behavior different from the previous scenarios, varying according to the profile of the class. This can reflect the influence of the teacher, as well as the content and the teaching methodology.

4.1.4. *Scenario 4: CAT X SAT*

Fig. 5 shows the comparative data between the means of the motivation and ability factors, separated by the samples using CAT and SAT. For a confidence level of 75%, the *anxiety* was a little higher for the CAT than the SAT. The estimated ability (θ), *expected success*, *interest* and *challenge* were statistically equivalent.

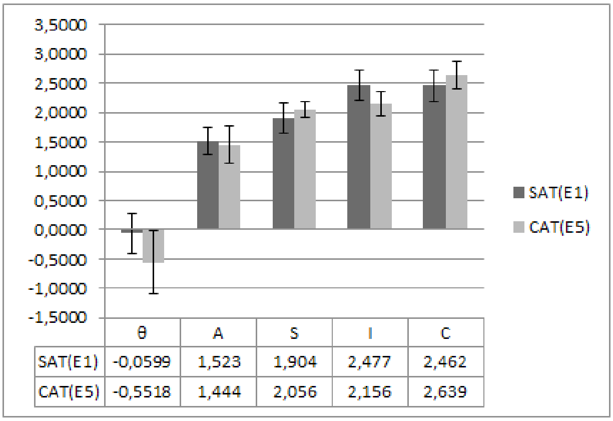
Considering the use of CAT, it was noted (Table 7) that students with higher *anxiety* expressed lower expected *success* rate. Likewise, the students with higher expected *success* rate (optimists) expressed less anxiety (moderate correlation of -0.615). Similarly, the more challenged students also felt more optimistic (a moderate correlation of 0.535). These facts were not observed in the execution of the SAT. There are also some correlations in common with SAT (Table 8): *interest* and *challenge*, result (number of right answers) and θ (estimated ability). There was no significant observed correlation between the motivation factors and the performance (*result* and θ).

Considering the Flow Theory (Csikszentmihalyi, 2000) and also the Inverted U Theory (Yerkes and Dodson, 1908), it is important to assess whether the greater adaptivity of CAT was advantageous for decreasing *anxiety*. Fig.s 6 and 7 shows scatter plots containing the mean difference between the estimated difficulty of the questions delivered before answering the motivation questionnaire and also the student's estimated ability for the SAT and CAT respectively, organized by class. Negative values on the x-axis represent that the set of questions delivered was on average more dif-



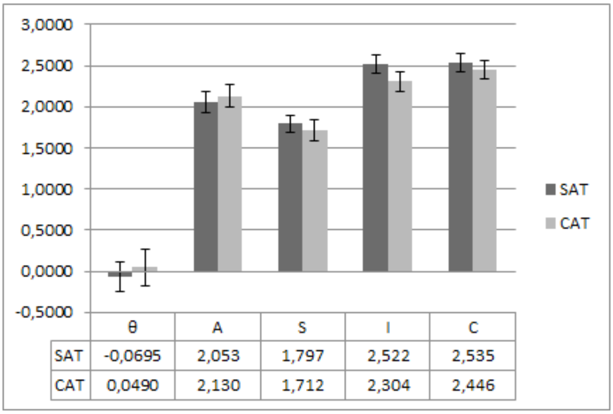
	θ	A	S	I	C
SAT(E6+E9)	-1,1882	2,729	2,036	3,014	3,000
CAT(E4)	0,2878	2,327	1,727	2,436	2,398

Fig. 3. Motivation factors of the Scenario 2.



	θ	A	S	I	C
SAT(E1)	-0,0599	1,523	1,904	2,477	2,462
CAT(E5)	-0,5518	1,444	2,056	2,156	2,639

Fig. 4. Motivation factors of the Scenario 3.



	θ	A	S	I	C
SAT	-0,0695	2,053	1,797	2,522	2,535
CAT	0,0490	2,130	1,712	2,304	2,446

Fig. 5. General motivation factors.

Table 7

Pearson's correlation in CAT

Factors	Anxiety	Expected Success	Interest	Challenge	Results	*θ*
Anxiety	1					
Expected Success	-0,615	1				
Interest	-0,200	0,457	1			
Challenge	-0,191	0,535	0,590	1		
Results	-0,144	0,376	0,257	0,200	1	
θ	-0,214	0,206	0,061	0,091	0,617	1

Table 8

Pearson's correlation in SAT

Factors	Anxiety	Expected Success	Interest	Challenge	Results	*θ*
Anxiety	1					
Expected Success	-0.271	1				
Interest	0.177	0.368	1			
Challenge	0.338	0.126	0.620	1		
Results	0.085	0.300	0.218	0.259	1	
θ	-0.077	0.219	0.221	0.109	0.519	1

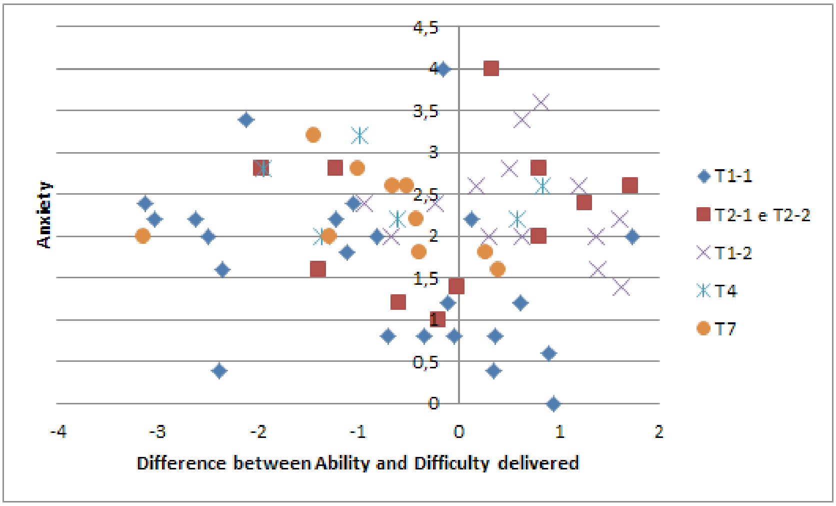


Fig. 6. Difficulty GAP x Anxiety (SAT).

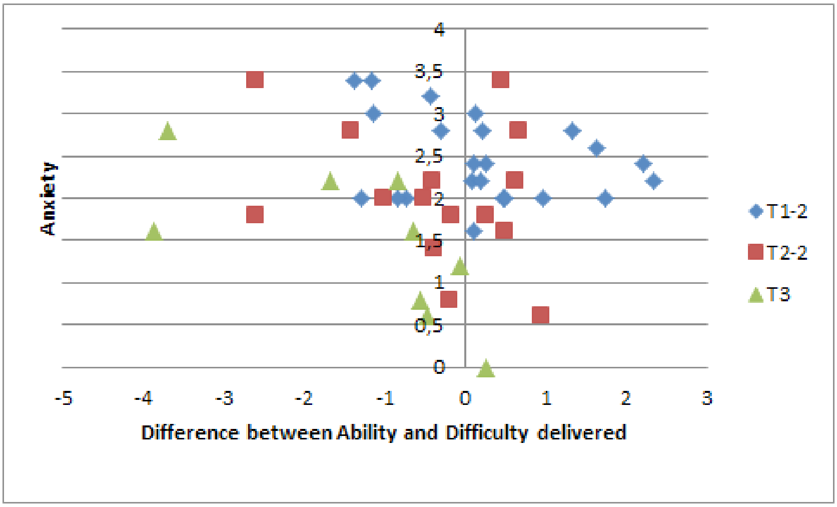


Fig. 7. Difficulty GAP x Anxiety (CAT).

ficult than the estimated ability. Equivalently, positive values mean easier sets than the estimated ability, while values close to the y-axis represent sets of questions more suited to the student.

In general, students in the same class were clustered on the graph, demonstrating a strong indicator that students reactions to an activity may reflect the influence of the teacher, as well as the content and teaching methodology applied. In addition, linear regression lines were calculated for each class and, except for classes T2-1 and T2-2 (Table 9), all others showed decreasing behavior as can be seen in negative angular coefficient B1 (Tables 9 and 10). This indicates that the more difficult is the set of activities given, taking as reference the student's ability, the greater is her *anxiety* tendency. A possible justification for the exception is that classes T2-1 and T2-2 are Computer Engineering classes, which in addition to being of a distinct program (the others are from Information Systems), had low participation (three collections of only three, one and three participants for T2-1 and a collection of four participants for T2-2), and because of this, low sampling may have influenced the observed behavior.

Table 9

Regression line coefficients in SAT

Class	B1	B0
T1-1	-0,2670	1,4101
T2-1 e T2-2	0,1223	2,2416
T1-2	-0,1448	2,4444
T4	-0,0847	2,4509
T7	-0,1460	2,1403

Table 10

Regression line coefficients in CAT

Class	B1	B0
T1-2	-0,1482	2,4842
T2-2	-0,2274	1,9507
T3	-0,3936	0,9372

Table 11

Difficulty evolution summary

Code	Topic	Num. items	Stabilized	Num. Observations
A1	Introduction	10	20	48
A2	Physical layer	5	10	31
A3	Data link layer	16	No	25
A4	Transport layer	15	No	12
A5	Application layer	8	19	30
A6	Network layer	15	No	20

4.2. *Evolution of the Estimated Value for the Difficulty of an Item*

In this section, we discuss the results of the evolution of the estimated values of difficulty of various questions by subject. The objective of this analysis is to find a reference for the number of samples (definition of sample is available in Section 3.6) necessary to well-infer the difficulty of an item, using the IRT, in a classroom context, which produces only a few tens of samples per semester.

Briefly, data related to from this part of the study are described in Table 11. Considering the methodology adopted, which asked students to answer at least five questions, there are indications that the larger the number of questions in the subject, the more samples are needed to calibrate. Possibly, if more students who participated in this research had fully explored the item bank, it would have been better to calibrate it. But, the data collected indicate that, to have a reasonably stable calibration, we need just a number of participants a little more than twice the number of items about a specific topic. In the context chosen to apply the results of this study, it is simple and scalable to obtain more samples (there are only a few dozen per semester). So, it is not interesting to postpone the use of IRT information only to seek to improve its accuracy.

4.3. *Discussion of Results*

Considering the four scenarios and the Pearson's correlation analyzes results, it was unexpected that CAT model provides more anxiety than SAT, since CAT provides more adaptability. One hypothesis that emerges with these data is that some phenomenon may have triggered *anxiety* in the execution of CAT, and made the students have low expected *success* in their performance. However, those who took the activity as a *challenge* overcame pessimism and, for a few cases, dispelled *anxiety* (weak but negative correlation of -0.191).

An assumption for the cause of anxiety in students using CAT is related to characteristics of the method. In CAT, when starting a test, the student receives the item with the closest difficulty to the average. If she selects the correct answer, the algorithm will estimate for her an ability that tends to infinity, selecting the item with the estimated highest difficulty level from the question bank. In a different way, the SAT starts with the item belonging to the stratum of lower level of difficulty and the algorithm delivers an item of the next stratum in case of hit, which constitutes a behavior of gradual increasing difficulty. Thus, the possibility of abruptly growing the level of difficulty in CAT may have caused a premature anxiety in the students, since the initial moments of the activity.

So, the adaptability provides more anxiety on students? No, as can be seen in scatter plots of Fig.s 6 and 7, there is evidence that a greater adaptability can positively influence the student's *anxiety*. Since the negative angular coefficient means that greater differences between difficulty of an item and examine ability reflect to greater anxiety. Considering these results and the authors' observations, it can be noted that adaptability could be improved if there were more items in the item bank, dispersed in different difficulties, especially at the most elementary levels. In this way, the system can compose the activities for a low ability student in a more appropriate way for her. However, the addition of many items can make it difficult to calibrate the items, as seen in Section 4.2.

5. Conclusions and Future Work

This work presented a comparative study between a Stratified Adaptive Test (SAT), based on the tacit knowledge of the teacher, and an IRT-based Computerized Adaptive Test (CAT), from the point of view of students' motivation.

The main contribution of this work is the proposed improvements to a teaching support tool. This tool aspires to engage students and increase their motivation which consequently lower their anxiety. Additionally, this tool would also appropriately foster a challenging and interesting learning environment. In addition, this tool can have characteristics of promoting learning respecting the inherent pace of each student, through adaptability. This tool could also function to monitor student development, to identify strengths and weaknesses, and provide formative assessments. Thus, it could be applied both in face-to-face teaching or in distance learning.

Some factors may limit the validity of the results of this study:

- There was a small sample population. That's because according to the country's laws, all participation must be voluntary.
- Technical problems prevented three data collections (E7, E8 and E10).
- Some students participated in more than one sample, since in the T1-1, T1-2, T2,1 and T2-2 classes the system was used in 3 to 4 different courses. Thus, there were students who used both CAT and SAT, and no study was done to

distinguish the samples that only took the test once, from those which repeated the tests more often.

To achieve the objectives, the two adaptive models were implemented in ATES.

Data analysis indicate that the levels of *anxiety* expressed in the use of SAT were lower than those expressed in the use of CAT. Besides, SAT is a simpler method to be implemented. The hypothesis that arises to explain a higher *anxiety* is that CAT can change to very difficult items at an early stage of the activity, scaring the student. However, it has been found that adaptability can positively influence anxiety levels, since the tendency of students who have been given more difficult questions than their ability is to declare themselves more *anxious*.

Considering the hypotheses raised in the interpretation of these data, the use of a hybrid model may be more advantageous. The implementation of a SAT in which the strata are composed initially based on the tacit knowledge of the teacher and later, through an IRT calibration, after a reasonably stable calibration (as seen in Section 4.2), combines the presentation of questions with a slightly increasing difficulty. This approach also has the accuracy and better adaptability of the IRT, collaborating for a system that optimizes student's levels of *anxiety*.

Therefore, it is left for future work to verify this hypothesis by implementing and testing a SAT in which the tests are formed from the proper calibration of the items, based on IRT and applied to face-to-face teaching.

Another interesting suggestion for future work is the characterization of the profile of the student, with regard to her attitude while doing an activity. This study revealed that there were different reactions, but similar reactions among students from the same class. So, we have the following research question to solve: what makes a class have more or less anxious reactions to the same challenge?

And finally, other future work is the development of an adaptive system that exploits Spacing Efect, estimating the student's ability and periodically suggesting to her the resolution of exercises that she has already studied, in order to strengthen learning and long-term memory.

Acknowledgments

The authors would like to thank the Brazilian Federal Agency for the Improvement of Higher Education (CAPES) for the financial support given to this project and Mr. Alexandre Teixeira for the contributions.

References

Araujo, R., Zorzo, A., Nunes, D., Matos, E., Steinmacher, I., Leite, J., Correia, R., Martins, S. (2017). *Referenciais de Formação para os Cursos de Graduação em Computação. Sociedade Brasileira de Computação (SBC)*. ISBN 978-85-7669-424-3.

Behar, P.A. (2009). *Modelos pedagógicos em educação a distância*. Artmed Editora, Porto Alegre.

Betz, N.E., Weiss, D.J. (1976). *Psychological Effects of Immediate Knowledge of Results and Adaptive Ability Testing*. Research Report 76-4. http://files.eric.ed.gov/fulltext/ED129863.pdf

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee´s ability. *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Chiu, L.-H., Henry, L.L. (1990). Development and validation of the mathematics anxiety scale for children. *Measurement and Evaluation in Counseling and Development*, 23(3), 121–127.

Couto, G., Primi, R. (2011). Teoria de resposta ao item (*TRI*): Conceitos elementares dos modelos para itens dicotômicos. *Boletim de Psicologia*, 61(134), 1–15.

Csikszentmihalyi, M. (2000). *Beyond Boredom and Anxiety*: Experiencing Flow in Work and Play. San Francisco: Jossey-Bass.

De Vos, T. (2010). *Manual Tempotoets Automatiseren*. Amsterdam: Boom Testuitgevers.

Ebbinghaus, H. (1885). *Über das Gedächtnis. Untersuchungen zur Experimentellen Psychologie* [Memory: A Contribution to Experimental Psychology]. Duncker & Humblot, Leipzig.

Eggen, T.J.H.M., Verschoor, A.J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, 30(5), 379–393.

Eklöf, H. (2010). Student motivation and effort in the swedish timss advanced field study. In: *Proceedings of the 4th IEA International Research Conference.* Gothenburg.

Finley, J.R., Benjamin, A.S., Hays, M.J., Bjork, R.A., Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, 64(4), 289–298.

Freund, P.A., Holling, H. (2011). Who wants to take an intelligence test? personality and achievement motivation in the context of ability testing. *Personality and Individual Differences*, 50(5), 723–728.

González-Sacristán, C., Molins-Ruano, P., Díez, F., Rodríguez, P., Sacha, G.M. (2013). Computer-assisted assessment with item classification for programming skills. In: *Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality*. ACM, New York, NY, USA, 111–117.

Hambleton, R., Swaminathan, H., Rogers, H. (1991). Fundamentals of Item Response Theory. *Measurement Methods for the Social Science*. SAGE Publications.

Hirose, H., Aizawa, Y. (2014). Automatically growing dually adaptive online IRT testing. In: *Proceedings of the IEEE International Conference on Teaching, Assessment and Learning for Engineering* (TALE), New Zealand, 528–533.

Hornke, L.F., Küppers, A., Etzel, S. (2000). Design and evaluation of an adaptive matrices test. *Diagnostica*, 46, 182–188.

Huang, S.X. (1996). A content-balanced adaptive testing. In: *Computer Aided Learning and Instruction in Science and Engineering. CALISCE-Computer Aided Learning and Instruction in Science and Engineering*. (3), 29–31.

Jansen, B.R., Louwerse, J., Straatemeier, M., der Ven, S.H.V., Klinkenberg, S., der Maas, H.L.V. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences*, 24, 190 – 197.

Jorba, J., Sanmarti, N. (2003). *Avaliação Como Apoio à Aprendizagem*. Artmed Editora, Porto Alegre.

Karadeniz, S. (2011). Effects of gender and test anxiety on student achievement in mobile based assessment. *Procedia – Social and Behavioral Sciences,* 15, 3173–3178.

Karpicke, J.D., Roediger, H.L. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968.

Landauer, T.K., Bjork, R.A. (1978). Optimum rehearsal patterns and name learning. In: M. Gruneberg, Morris EE, Sykes, RN (Ed.), *Practical Aspects of Memory*. 625–632.

Lilley, M., Barker, T., Britton, C. (2005). Learners' perceived level of difficulty of a computer-adaptive test: A case study. In: Costabile, M.F., Paternò, F. (Eds.), *Proceedings of the Human-Computer Interaction –* INTERACT 2005: IFIP TC13 International Conference, Springer Berlin Heidelberg, Berlin, Heidelberg, 1026–1029.

Lord, F. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 17(2), 181–194.

Lord, F. (1980). Applications of Item Response Theory to Practical Testing Problems. Erlbaum Associates.

MacIntyre, P.D. (2002). *Individual Differences and Instructed Language Learning*. Vol. 2. John Benjamins Publishing, Philadelphia.

Manseira, P.R.P., Misaghi, M. (2013). Proposta de ferramenta para uso abrangente de testes adaptativos computadorizados na educação a distância. In: *Anais do III CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO*, Ponta Grossa.

Meijer, R.R., Nering, M.L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23(3), 187–194.

Mukaka, M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3), 69–71.

Oppermann, R., Rashev, R., Kinshuk. (1997). Adaptability and adaptivity in learning systems. *Knowledge Transfer*, 2, 173–179.

Ortner, T., Weißkopf, E., Gerstenberg, F. (2013). Skilled but unaware of it: Cat undermines a test taker's metacognitive competence. *European Journal of Psychology of Education*, 28(1), 37–51.

Ortner, T.M., Weißkopf, E., Koch, T. (2014). I will probably fail: Higher ability students motivational experiences during adaptive achievement testing. *European Journal of Psychological Assessment*, 30, 48–56.

Pardos, Z.A., Gowda, S.M., Baker, R.S., Heffernan, N.T. (2012). The sum is greater than the parts: Ensembling models of student knowledge in educational software. *ACM SIGKDD Explorations Newsletter*, 13(2), 37–44.

Penk, C., Schipolowski, S. (2015). Is it all about value? bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences*, 42, 27–35.

Pintrich, P.R., Smith, D.A., García, T., McKeachie, W.J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (mslq). *Educational and Psychological Measurement*, 53(3), 801–813.

Piton-Gonçalves, J., Monzón, A.J.B., Aluísio, S.M. (2009). Métodos de avaliação informatizada que tratam o conhecimento parcial do aluno e geram provas individualizadas. In: *Proceedings of the Conference on Simpósio Brasileiro de Informática na Educação (SBIE)*, Sociedade Brasileira de Computação, Florianópolis, SC, Brazil.

Reeve, J. (2014). *Understanding Motivation and Emotion*, 6th Edition. John Wiley & Sons.

Ricarte, T.A.M. (2013). *Teste Adaptativo Computadorizado nas Avaliações Educacionais e Psicológicas*. Master's thesis, Universidade de São Paulo, São Paulo.

Rudner, L.M. (2012). *PARAM – Calibration Software for Logistic IRT Models*. `http://echo.edres.org:8080/irt/param/`

Silva, R., Direne, A. (2016). Sequenciamento adaptivo de exercícios baseado na correspondência entre a dificuldade da solução e o desempenho dinâmico do aprendiz. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, Uberlândia, p. 11.

Tesseroli, R., Direne, A., Pimentel, A., Spinosa, E., Melniski, L. (2016). Geração automática de avaliações utilizando o algoritmo abc para definição da próxima questão. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, Uberlândia, p. 407.

Thompson, N.A., Weiss, D.J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, 16(1).

Veerman, J., Straathof, M., Treffers, P.D., Van den Bergh, B., Ten Brink, L. (1997). *Dutch Manual for Perceived Competence Scale for Children*. Pearson, Amsterdam.

Vollmeyer, R., Rheinberg, F. (2006). Motivational effects on self-regulated learning with different tasks. *Educational Psychology Review*, 18(3), 239–253.

Weiss, D., Bock, R. (1983). *New Horizons in Testing: Latent trait Test Theory and Computerized Adaptive Testing*. Elsevier.

Yeh, D.D., Park, Y.S. (2015). Improving learning efficiency of factual knowledge in medical education. *Journal of Surgical Education*, 72, 882–889.

Yerkes, R.M., Dodson, J.D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5), 459–482.

S. Yoshioka has a BSc (2010) in Computer Science and MSc (2017) in Informatics from Pontifical Catholic University of Minas Gerais (PUC Minas), Belo Horizonte, Brazil. His research interests are focused on item response theory and computer adaptive test applied on education or enterprise environment for learning or performance evaluation.

L. Ishitani is a computer science Professor at Pontifical Catholic University of Minas Gerais (PUC Minas), Belo Horizonte, Brazil. She has a BSc (1990), a MSc (1993) and DSc (2003) in Computer Science from Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil. Her research interests are focused on computers in education, learning objects, serious games and interaction design. She works as a reviewer for several journals and conferences in the field of informatics in education.