International Journal of User-System Interaction 12(1) 2019, 1-20

SITAC – Innovative Computerized Adaptive Testing System

Servet Giafer, Cristina Giafer

S.C. Soft Business Union S.R.L.
6 Blv Regina Maria, sector 4, 040125, Bucharest, Romania

Abstract. The computerized adaptive testing is an approach of the differential assessment which adapts the questions that are asked to the candidate's ability level. Thus, the computer selects and displays the questions, then records and processes the candidate's answers. The items' selection is adaptive and it depends, on one hand, on the candidate's answers to the previous questions and on the statistical qualities of the items, on the other hand. As compared to the traditional testing methods, where all candidates receive the same items, the computerized adaptive testing is managing a higher percentage of items, with appropriate difficulty levels. The items' adaptive selection process results in higher levels of the scores' precision and shorter tests. *SITAC* - the Innovative Computerized Adaptive Testing System represents a complex and innovative software product based on a computerized adaptive testing platform through which the testing of the persons will be performed, for the purpose of estimating their abilities, based on the answer previously delivered.

Keywords: Computerized Adaptive Testing, Item Response Theory, Item Characteristic Curve, Logistic Model.

1. Introduction

The Computerized Adaptive Testing (CAT) is not a new concept, but it started to gain popularity along with technology development, which enabled easier implementations. At the international level, more and more organizations have grown conscious of the added-value that this testing manner may bring and choose to purchase such systems.

The CAT modern algorithms are based on concepts from the Item Response Theory (IRT) and from statistical theories such as the maximum plausibility and the Bayesian estimation. IRT represents the study of the tests and questions' assessment made on the basis of the assumptions regarding the mathematical relation between the ability of a candidate and the answers he/she gives to the questions. The adaptive tests based on the IRT contribute

to reducing the tests' duration, as the adaptive test offers the questions that are most relevant for each candidate.

Each examinee takes a unique test that is tailored to his or her ability level. Questions that have low information value about the test taker's proficiency are avoided such that test takers are challenged by test items at an appropriate level. They are not discouraged or annoyed by items that are far above or below their ability level. Because each test is unique to the examinee, it is more difficult to capture the entire pool of items. Doing so would require the careful collaboration of many examinees of varying ability levels. Less time is needed to administer CAT than fixed-item tests because fewer items are needed to achieve acceptable accuracy. Adaptive tests can result in a 50%–90% reduction in the number of items administered, with no decrease in measurement quality (Brown & Weiss, 1977).

The structure of this article is the following: The following section presents a detailed description of the concepts based on the IRT, followed by details on the structure of the CAT systems. After describing the *SITAC* project (principles, interfaces, results), the conclusions and future work section encourage the use and development of such systems for the purpose of improving them.

2. Item Response Theory

2.1 History

The Classical Test Theory (CTT) was developed in the 1920's (Bock, 1997) This theory is made up of multiple theories, such as the validity theory, the reliableness theory, the tests analysis theory, the items analysis theory, etc. Most of the practices were initially limited to psychological tests, being extended also to the education area later. However, a new testing theory, with a conceptual power higher than the classical one, was developed during the last 50 years. This is based rather on items than on the test scores and is known as the Item Response Theory (IRT).

While the basic concepts of the IRT were and remained simple, the mathematics which is the basis for it is, though, advanced enough, as compared to the one used on the CTT. Thus, the examination of some of these concepts was difficult enough, without the performance of a great number of

calculations in order to get useful information by using the informatics technology; the evolution of one influenced, significantly, the other one.

2.2 IRT vs. CTT

IRT presents a series of benefits in the psychological and educational tests field, as it offers more adaptable and effective methods, together with their analysis and grading, unlike the ones derived from the CTT. IRT sets up a series of relations between the items' properties and the operational features of the test which is made up of items. These relations can be validated for real tests having any duration. The IRT vision as to approaching the items or the small sets of similar items as interchangeable tests construction and grading units resulted in many innovations for the testing practice, especially when building item banks and adaptive tests. The first can significantly reduce the time and costs needed for producing a high-quality operational test. The second one, which comes either as computerized adaptive testing or in two testing stages, by using the `paper and pencil' form, enables the shortening of the testing time to half or one-third of the duration needed for a conventional test, with the same precision.

The capacity of withdrawing and replacing the items in one operational test without modifying the interpretation of the grading scale has the same importance for the long term testing and assessment programs. As the scores on the IRT scale are functions of the items' estimated parameters, the calculation of scores will absorb possible differences of characteristics (difficulty, discrimination, etc.) between the withdrawn items and their replacements. Thus, the necessity to identify new items, with the same level of difficulty and discrimination power as the old items or for the study equivalent to the test reviewed separately of their operational use is eliminated, as requested in the classical theory.

Another unique feature of the IRT is placing the items and candidates on the same scale. Response models on which the IRT models are based enable the analyst to specify the probability that a candidate with a certain score level shall respond correctly to a certain item. In CTT, the raw score of the candidate is the sum of all scores received for the test items. In IRT, we are mainly interested in whether a candidate answered correctly or not to each item and not in the raw scores of the test. This is due to the basic concepts of the IRT, which are based upon individual elements of a test, rather than on the answers.

Thus, it is clear that, in CTT, the test and the entire population of candidates are considered together, and the statistics are calculated based on this assumption. In IRT, we highlight the individual item and the individual candidate.

2.3 Item Characteristic Curve

In case of a typical test based on items, on any value of the ability, the probability of a correct answer $P(\theta)$ will be reduced for the candidates with low ability and will be great for the candidates with high ability. The probability of a correct answer is almost 0 for the lowest levels of ability. This increases until it gets close to the value 1 for the highest levels of ability. This S-shaped curve (Figure 1) describes the relationships between the probability of a correct answer and the ability scale. In IRT, this is named the item characteristic curve (ICC) or the item response function (IRF). Taking into account the fact that each item of the test has a different difficulty, each item will have its own ICC.

The result of the graphic representation of an ability function $P(\theta)$ is an S-shaped curve (Figure 1).

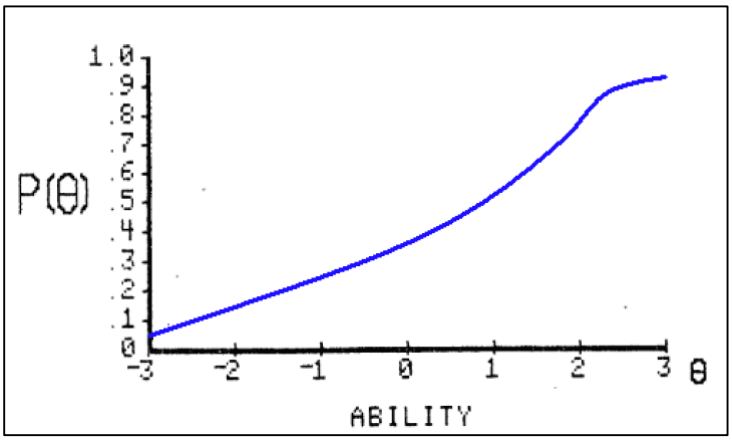

Figure 1: A general characteristic curve (Baker, 2001)

ICC is the basic block of the item response theory; all of the other concepts of the theory are dependent on this curve. There are two technical features of the ICC used for describing that. The first is item difficulty. In IRT, the difficulty of an item describes where the item stands on the ability scale. For

example, an easy item is adequate for persons with reduced ability, while a difficult item is adequate for persons showing a high ability level; thus, we may say that difficulty is a localization coefficient. The second technical feature is discrimination, which describes how well an item can make the difference between the candidates with abilities below the position of the item and those having the level of abilities above its position. This feature reflects, in fact, how steep the ICC slope is. The more plane the curve is, the less will the item discriminate, as the probability of a correct answer on lower levels of ability is almost the same with the one for the higher levels of ability.

Using these two descriptors, the general shape of the ICC can be described. These descriptors are used, also, for discussing the technical features of an item.

In Figure 2, all characteristic curves are presented in the same chart. All these features the same level of discrimination but differ as regards the difficulty. The green curve represents an easy item, as the probability of a correct answer is great for candidates with the reduced ability and it gets close to 1 for the candidates with a high level of ability. The red curve represents a medium difficulty item, as the probability of a correct answer is reduced on candidates with reduced abilities, i.e., around 0.5 in the middle of the ability scale, and almost 1 for high levels of ability. The blue curve represents a difficult item for which the probability of a correct answer is low for persons with reduced abilities and will increase when higher levels of ability are attained.

The discrimination concept is illustrated in Figure 3. The figure contains three ICC with the same level of difficulty, but different levels of discrimination. The magenta curve features a high level of discrimination, as its slope is steep enough in the middle, where the probability of a correct answer is changing very fast, as the ability level increases. Just a short distance to the left of the middle of the curve, the probability of correct response is much less than 0.5, and a short distance to the right the probability is much greater than 0.5. The blue curve represents an item with a moderate level of discrimination. The slope of this curve is smaller as compared to the previous curve, as the level of ability increases. However, the probability of a correct answer is almost 0 for candidates with the lowest levels of ability and almost 1 for candidates with the highest levels of ability. The green curve represents an item with weak discrimination. This features a very short slope,

and the probability of a correct answer is changing slowly along with the displayed ability range.

Even on lower levels of ability, the probability to get a correct answer is high enough and it will only increase a little when higher levels of ability are reached (although the image shows an ability range that varies between -3 and +3, this may, theoretically, vary between $-\infty$ and $+\infty$).

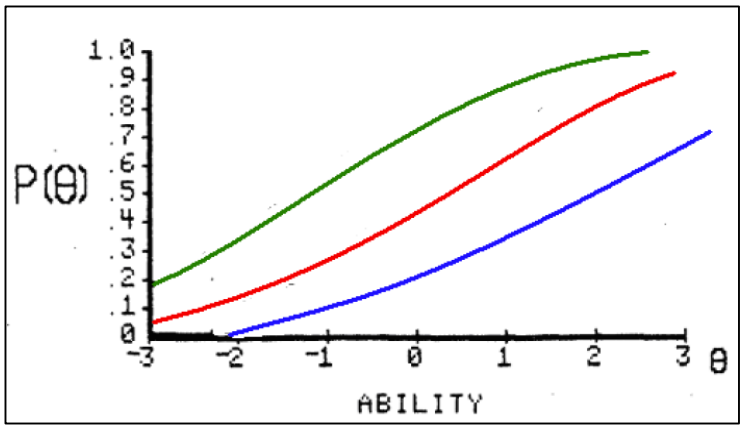

Figure 2: Three characteristic curves with the same discrimination level, but featuring different difficulty levels (Baker, 2001)

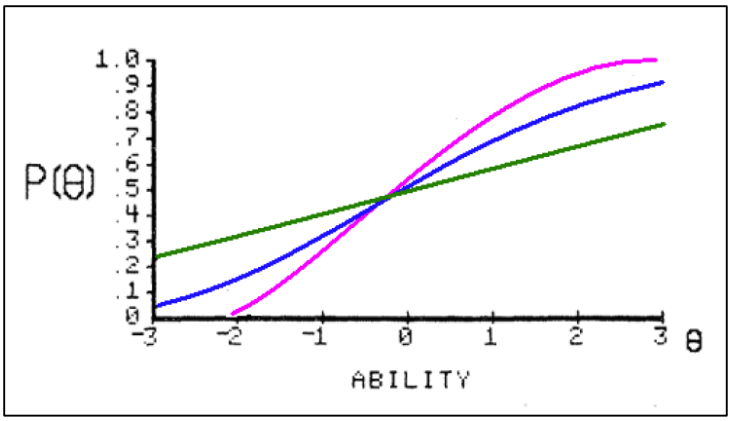

Figure 3: Three characteristic curves with the same difficulty level, but featuring different discrimination levels (Baker, 2001)

3. Characteristics of the CAT systems

3.1 Basic component parts

When IRT is used for processing the test data, the items' calibration is a crucial component of the testing's operation. Item's calibration refers to the IRT models' adjustment procedure using the data made up of the answers collected from a sample group of candidates and the estimation of parameters using these data.

Modern assessments are based on great banks of items. Each item in the bank must be calibrated before being used in operational tests, and the accuracy of the calibrated items' parameters affects directly the validity and reliableness of the test, which reflects upon the candidate's assessment, the equivalence and the differential analysis of the item's functioning, etc.

A CAT system is, basically, made up of the following parts:

1. Setting up the initial θ value;
2. Methods of item selection;
3. Methods of intermediate and final estimation of the θ value;
4. Completion criteria.

An initial θ value is necessary for a candidate at the beginning of the test, when no preliminary information about the candidate is available. A simple option is the use of the anticipated average distribution of the abilities as the initial value for all tested candidates. It is also possible to include a random value, in order to avoid offering certain similar initial items to the candidates.

The methods of item selection represent the most important part of the CAT system. These must not only serve the purpose of optimizing the statistical efficiency as regards the estimation of the candidates' ability parameters but also to fulfill more non-statistical constraints (e.g., balancing the content, number of words) and to control the degree of exposure of every item.

The estimation of the ability can be made through a variety of methods developed for the IRT, the most used one being presented as follows:

The component part regarding the completion criteria aims both the fixed length and the variable length CAT, with different rules, such as completing the test once a satisfactory level has been obtained for the standard error mean (SEM).

3.2 Logistic models

IRT uses a variety of probabilistic models in order to shape up the probability of a correct answer if the item is assessed from the dichotomic point of view or some different levels of the answer if the item features more than two possible assessment levels (polychotomic). These probabilities depend on the item's parameter and the candidate. When items are assessed from dichotomic point of view, i.e., by asking multiple answers questions, the most common IRT models used are the logistic models with one, two or three parameters (1PL, 2PL, 3PL). The probability of a correct answer for an item j assigned to a candidate regarding the ability θ is shaped up by the following item response functions (IRF):

- The logistic model featuring one parameter:

$$P_j(\theta) = 1/(1 + e^{-1(\theta - b_j)})$$

- The logistic model featuring two parameters:

$$P_j(\theta) = 1/(1 + e^{-a_j(\theta - b_j)})$$

- The logistic model featuring three parameters:

$$P_j(\theta) = c_j + (1 - c_j)\left[1 \,/\, 1 + e^{-a_j[\theta - b_j]}\right]$$

Where:

a_j = discrimination of item j;
b_j = difficulty of item j;
c_j = the guess factor for item j;

All these parameters of the item which vary according to the individual items describe the features of each item. The 1PL model is the simplest of all three but has the highest assumption degree: all items have the same discrimination power and the chances of guessing the correct answer are not taken into consideration. The 2PL model does not consider guessing the correct answer, but it supposes a various discrimination power, shaped up by means of parameter *a.* The 3PL model includes all the three parameters, which may delimit a more complete profile of an item. The purpose of

calibrating the items is to assess these parameters by using statistic algorithms for a data sample made up of the candidates' answers.

In case the answers related to an item can be assessed with more than two levels, such as the short answers with partial credit, a series of IRT polychotomic models can be used for shaping up the answer related data, such as the gradual answer model (GAM; Samejima, 1969), the partial credit model (PCM; Masters, 1982), the generalized partial credit model (G-PCM; Muraki, 1992), the rating scale model (RSM; Andrich, 1978), or the nominal answer model (NAM: Bock, 1972).

In the IRT, both the candidate's parameters and the item's parameters are invariable in relation with the sample. This means that, if a different set of items are administered to a candidate, the assessment of the ability parameter should generate the same value, excluding the random perturbation. Also, if a different group of candidates would get the item, the assessment of the items' parameters should, also, generate the same values, excluding the random perturbation. This feature of invariant is related to the sample sets up the basis of the adaptive selection of the item within the CAT.

3.3 The Fisher Information

The IRT, the standard error measurement (SEM) is not constant anymore for different levels of ability. Otherwise, the Fisher information, a classic statistical indicator, has been introduced in the IRT in order to supply the lower limit of the SEM's square on each θ level.

In case of the dichotomic IRT models, the Fisher information regarding the θ ability can be reduced to the form:

$$I(\theta) = \sum_{j=1}^{J} \frac{\left[P_j'(\theta)^2\right]}{P_j(\theta)\left[1 - P_j(\theta)\right]}$$

Where:

j = 1, 2 …, J – the J items to which the candidate answered;
$P_j(\theta)$ = the item response function formulated for the logistic models 1PL, 2PL, 3PL.

The Fisher information has a crucial role in selecting the items within the CAT. The most well-known method of selecting the item in the CAT, the maximum Fisher information method (Lord, 1980), selects the item which

maximizes the Fisher information for the estimation of θ. This method maximizes directly the SEM for θ on the asymptotic level. The Fisher informational matrix for the parameters' vector of the item is also important during the items calibration stage for the optimum sampling of the candidates.

4. SITAC – Innovative Computerized Adaptive Testing System

4.1 Context

The classical test theory is intended for developing conventional tests which use a fixed set of questions/items selected on the basis of the data related to a group of respondents. Even though this theory governed a multitude of testing instruments in time, it features a series of drawbacks of which the most important is represented by the low precision in measuring the candidates' abilities, as, many times, these do not receive questions related to the abilities they have.

Along with the development of the information technology and the refinement of the mathematical models, the ways of testing and assessing have progressed, as well. Thus, a new generation of tests started to gain ground as much as possible, worldwide, due to a new testing paradigm: the computerized adaptive testing.

4.2 Description

SITAC is implementing these paradigms for the purpose of increasing the contribution of the information technology and communications sector to the development of the Romanian economy. The innovative character of the project consists of the adaptive testing technology which supposes the elaboration of tests related to the level of the candidate which takes them. Thus, the proposed system performs the automatic adjustment of the examination according to the tested candidate's level of knowledge: in case the candidate answers correctly to the question, he/she will be asked more difficult questions successively, and, in case this answers incorrectly, he/she will then be asked easier questions, successively.

The system performs, thus, a more accurate assessment of the candidate, as compared to reality, taking into account that this will be tested according to the level of knowledge he/she owns, with no need to test lower or higher levels of his/her ability. In other words, it is not necessary to ask a candidate with a high level of knowledge very easy questions, as it is obvious that he/she has more advanced knowledge and knows the answer to those questions.

From the point of view of the measurement, two main objectives of the testing can be distinguished:

1. Estimation - the purpose is to get a valuation of the expertise, ability or performance of a person in a field that is well defined on a one-dimensional scale. Traditionally, CAT is designed for reaching this objective as fast and as precisely as possible;
2. Classification - the purpose is to determine which of the levels limited by expertise or performance classes a person belongs to.

In this case, one or more demarcation points shall be set up on the ability scale in order to decide the category a person belongs to. A precise estimation of the person's ability is not so important, but rather the classification within a certain category.

The basic principle of adaptive testing is simple: avoiding asking the tested candidates questions that are too difficult or too easy. Because we are sure enough (but not completely sure) that those candidates with high levels of ability will answer the easy items correctly and that those candidates lacking some skills will blunder into difficult questions and that we cannot jump to any useful conclusions following these answers. Much more useful is to ask the questions which pose challenges to the candidates, without overwhelming them. The correct identification and the subsequent asking of these questions represent the purpose of this adaptive test.

The adaptive test consists of 2 main steps, the questions selection, and the score estimation, respectively.

The first step determines the most appropriate question (or collection of questions) to be asked, taking into account what we know about the candidate's level of performance. The selection is made from a pool of questions containing more questions than the test given to the candidate.

The second step uses the answers to the questions that were previously asked, in order to refine the score of the candidate or the performance assessment. This allows the following questions to be more appropriate. This cycle is continued until either a certain number of questions were asked, or a certain precision was reached in estimating the score.

A diagram of the process is presented in Figure 4.

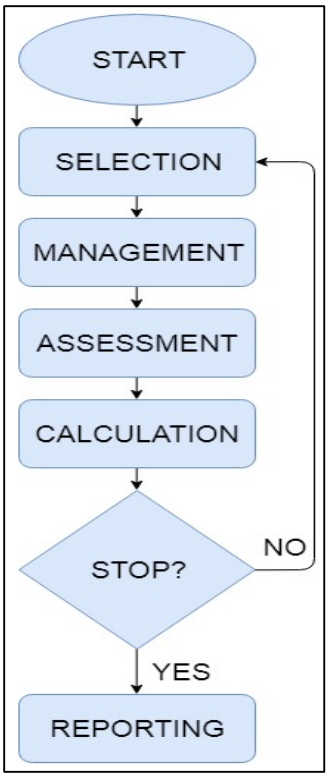

Figure 4: CAT flowchart

Start. After each item is given, a selection procedure for the next item is implemented. An item will be chosen from the bank of items, which is in accordance with the answers given by the candidate up until that moment: thus, adapted or created on the level of the tested candidate's quality.

Selection. The item selection procedure is responsible for increasing the guaranteed efficiency of the CAT. This is based upon the information concept the basic idea of which is that the item which is supposed to offer the best informational value for the candidate with the ability θ demonstrated by that moment will be given.

Management and assessment. These stages of the CAT algorithm are clear enough: item is presented, the candidate answers the item, the answer is assessed.

Calculating the score. During the calculation stage, the candidates' scores are processed. The statistic procedures determine, based on the items' score, the assessment of the ability θ and an indicator regarding the precision of this estimation.

Stop. After the management of each item, a decision is made whether to select another item or the testing can be ended. The criteria for completing the testing are:

1. The accuracy of estimating the ability θ;
2. The accuracy of the decision to classify the candidate;
3. The maximum (and possibly, the minimum) available time for testing or the number of items which can be managed.

Reporting. Possible reports:

1. Report on the estimated θ ability;
2. Simple graphic report on the category in which the candidate is classified.

The application *SITAC* contains five testing areas amounting to over 4200 calibrated items, respectively:

1. Determining primary psychological factors at the end of secondary school;
2. The psychological establishment of nodal points in choosing the right career;
3. The psychological highlighting of positive aspects and improving negative aspects targeting individual performance;
4. Accountancy;
5. Civil servants.

The first three areas belong to the psychological field. They consist of both right/wrong answer questions to determine the cognitive abilities and no correct answer questions to determine the personality traits and areas of interest. These tests were designed to be used especially in schools their main purpose being advising students in their career choices. Furthermore, they can be used very well in companies and public institutions for employee assessments.

The accountancy and the civil servants test only contain right/wrong answer questions and their purpose is to determine the ability level in these areas.

In addition to these predefined testing areas, *SITAC* allows the user to define an unlimited number of them. New areas can be defined by administrators by inputting certain specific information and adding new calibrated items.

In order to create a new testing area, the user has to fill in the required information (name, language, keywords, etc.). The next step is to add new calibrated items by entering the questions, answer options, calibration parameters and response time. It also allows the addition of images both in the question and in the answer options (Figure 5).

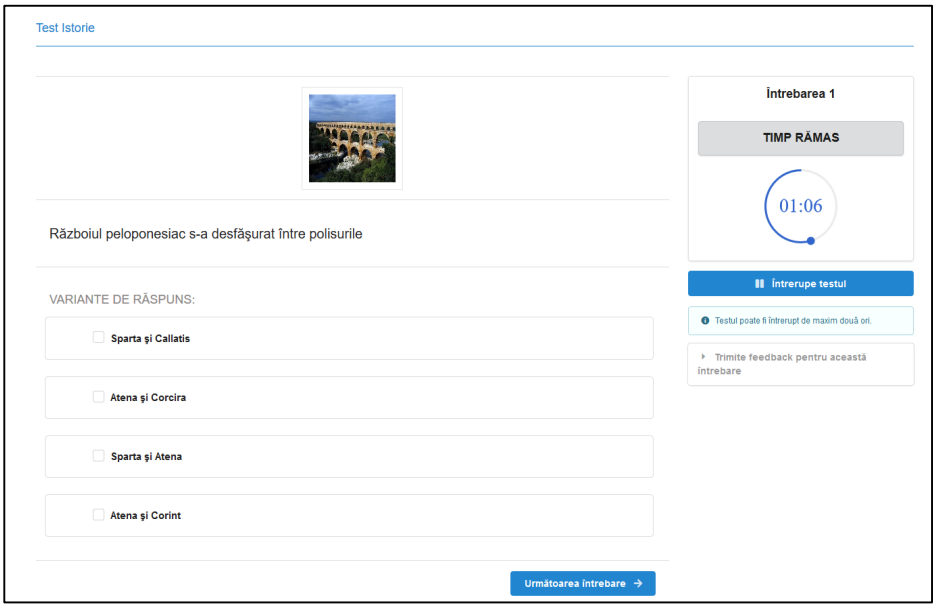

Figure 5 – An example of a multiple answer question having an image in its description

The system also allows for the management of invitations to the testing platform. These can be defined by administrators, and the target groups will be notified by email.

The test starts by providing the candidate a question of medium level difficulty. The algorithm will adjust the following question considering the candidate profile based on the previous responses. This process will be repeated until a stop criterion is verified.

If the candidate does not respond to an item in the assigned time, the

answer to that item will be considered wrong. The candidate has the possibility to pause a test two times and resuming the test where it was left.

5. User Interface / Results

The viewing of the results of the test differs according to the role of the user accessing them. Thus, the candidate may view the result as a report, both textual and graphical, while the examiner is supplied other additional information such as the personal data of the candidate and the answers to each item.

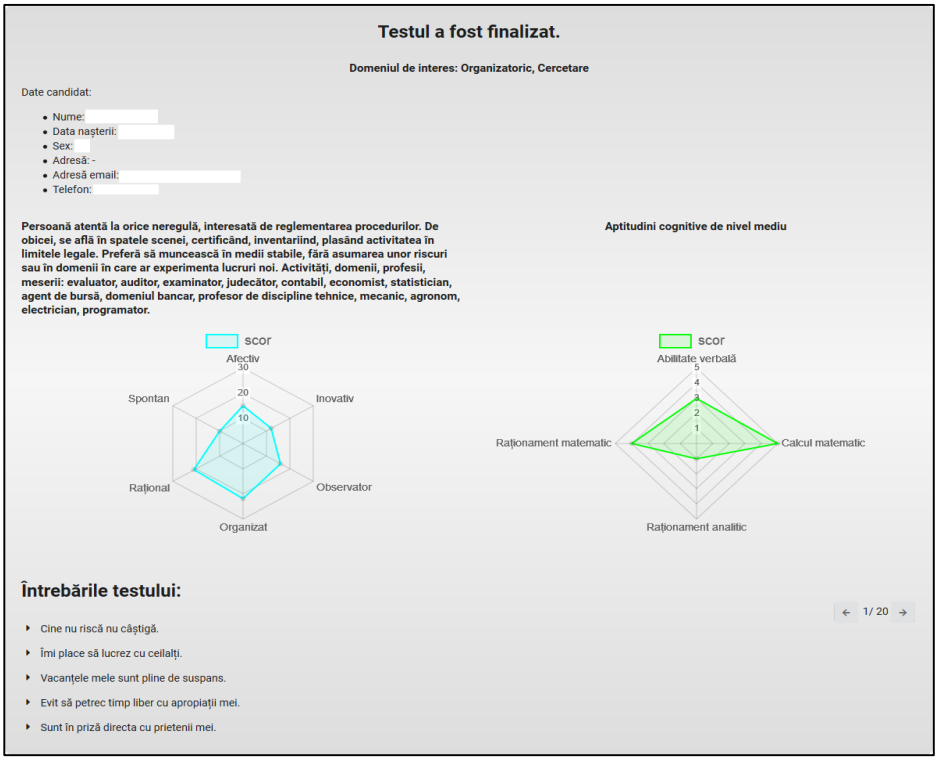

Figure 6: Test result example - examiner interface (psychological test)

Figure 6 shows the result of a person in a psychological test. It is represented both in textual and graphic form and measures both the

personality traits of the person in question and the cognitive skills. For an easier interpretation, the dimensions are represented using radar charts.

For tests that determine the ability of a candidate in a particular field, the result displays a report containing the total number of questions, the number of the right answers, the test duration, the ability level, and an ability scale. The report contains also a graphic in order to compare the candidate's results with the results of other candidates (Figure 7).

The candidate can view his/her score at any test he/she took in the past in order to compare the results and to become more aware of his/her evolution.

Moreover, the candidate can send feedback at any question he/she wants in case he/she notices something that has to be reported.

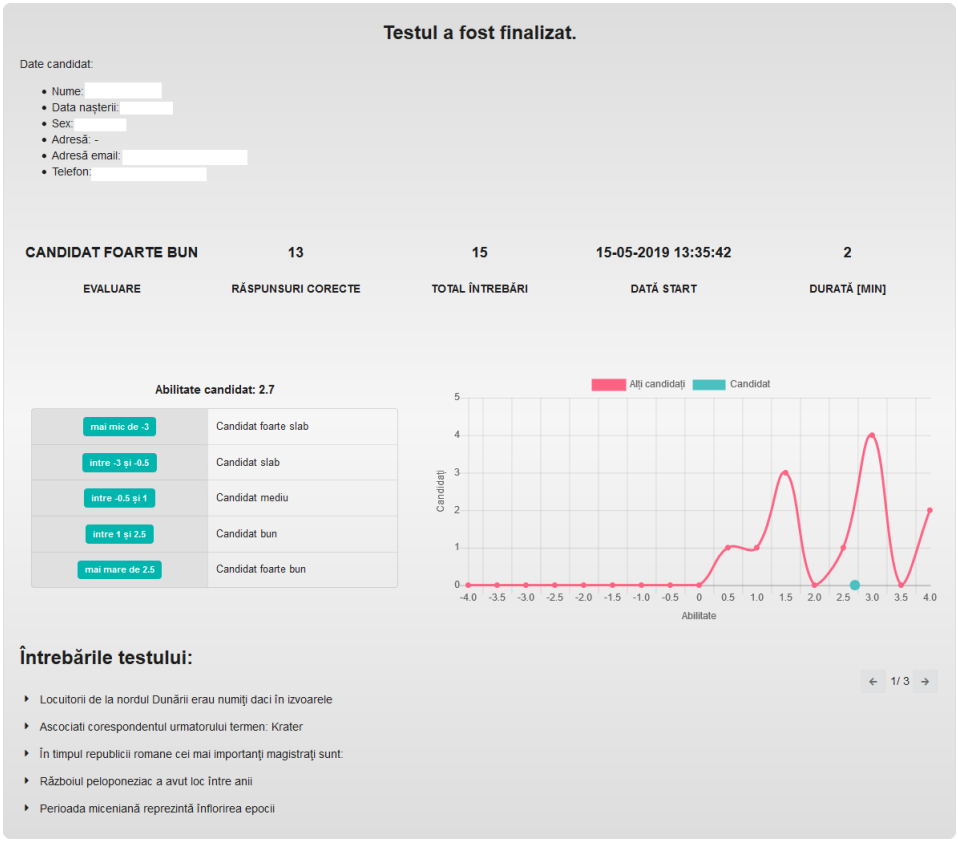

Figure 7: Test result example - examiner interface (custom history test)

6. Conclusions and Future Work

By eliminating the management of the items with inappropriate difficulty, the computerized adaptive testing may reduce the testing time, increase the precision of the measurements and reduce the error of measurement appearing due to boredom, frustration or guessing.

In the computerized adaptive testing systems, the level of ability of a candidate is re-assessed after each item resolved on the basis of the answers given for all previous items and the testing is completed when a certain criterion for measuring precision is met.

The computerized adaptive testing is based upon the Item Response Theory. This describes a set of probabilistic models wherein a set of parameters defining an item (difficulty, discrimination, guessing) interact with the level of ability a candidate owns in order to determine the probability of a correct answer for that item.

Within the item response theory, the ability and difficulty of the item are placed on the same scale.

If the ability of the candidate is relatively high as related to the item difficulty, the probability of a correct answer for that very item is high.

Otherwise, if the ability is reduced in relation to the level of difficulty, the probability of a correct answer will decrease and that of an incorrect answer will increase.

IRT models benefit from some features which make them useful for the CAT systems. One of these is the parameters invariance concept which states that the items' parameters are independent of the group of candidates for whom these are calibrated, and that the assessment of the ability is independent of the particular items a candidate receives.

The effects of such a system are obvious both from the point of view of the organization responsible for the testing and of the candidates sustaining it. For the first ones, the use of *SITAC* will massively reduce the costs with the development of tests as well as the costs with other administrative elements involved by such a process (e.g., costs with organizing a test) and will offer a better support for selecting the appropriate candidates, increasing the efficiency and productivity of the company. For the candidates, using *SITAC* will reduce the time spent during testing, will improve the testing experience and will also offer an accurate reflection of the level of knowledge on the basis of which he/she can further develop skills.

Although the adaptive testing systems have made significant progress,

these also featured a series of limitations. A major limitation is represented by the lack of possibility to review, as the CAT systems and, in particular, *SITAC,* do not allow a return to the items having already been administered. This return would be necessary in case of failed starts due to the candidates' severe anxiety. Also, the *SITAC* is designed to select the best items from the bank, which means that these items will become overexposed, an aspect which could be improved by implementing an exposure rate control algorithm.

CAT requires careful item calibration. This, in turn, requires that extensive data be collected on a large item pool. The development of a sufficiently large item pool is one of the biggest constraints to the widespread use of CAT.

Another aspect not to be neglected is the one related to public relations. Due to the complexity and the significant difference in relation with the familiarity of the traditional testing paradigm, an organization must make more efforts in the area of public relations in order to explain CAT and the reasons for using it.

However, it must be taken into account that the CAT has also a major social impact over the tested candidates. Considering the experiments that were performed in a large scale, it was demonstrated that the adaptive algorithm for construction of questions has a positive impact on student willingness to use the system (Papoušek & Pelánek, 2015). Nevertheless, in the artificial intelligence in the education community, this aspect is worth attention since it is usually not studied or taken into account.

Acknowledgments

The solution presented hereby was financed through the project SITAC "SISTEM INOVATIV DE TESTARE ADAPTIVĂ COMPUTERIZATĂ"" (INOVATIVE COMPUTERIZED ADAPTIVE TESTING SYSTEM), within the Competitiveness Operational Program 2014 - 2020, Priority axis 2 - Information Technology and communication (IT&C) for a competitive digital economy, Action 2.2.1 - Supporting the increase of the added value generated by the IT&C sector and of the innovation in the domain by developing clusters, Funding agreement no. 15/16.06.2017, Code MySMIS 115933.

References

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika

43(4), 561–573.

Baker, Frank B. (2001). The Basics of Item Response Theory.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 37(1), 29–51.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika 47(2), 149–174.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. Applied psychological measurement 16(2), 159–176.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement.

Lord, F. M. (1980). Applications of item response to theory to practical testing problems.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer (Ed.), Computer adaptive testing: A primer 171(199). Hillsdale, NJ: Lawrence Erlbaum.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). Computerized adaptive testing: From inquiry to operation. American Psychological Association.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. Proceedings of the 27th annual meeting of the Military Testing Association, 973-977.

Van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. Journal of Educational and Behavioral Statistics, 29 (3), 273-291.

Bock R.D. (1997). A brief history of item theory response. Educational Measurement: Issues and Practice, 6 (4), 21–33.

Peng Lu & Xiao Cong (2016). The Research on Computerized Adaptive Testing. Journal of Physics: Conference Series 710 012029.

van der Linden Wim J. (2008). Some New Developments in Adaptive Testing Technology. Journal of Psychology, 216, 3-11.

Lu Peng, Dongdai Zhou, Shaochun Zhou & Xiao Cong (2012). Design and Implementation of Computerized Adaptive Testing System for Multi-Terminals Modern Educational Technology, 22 88-92.

van der Linden W.J. (2008). Using response times for item selection in adaptive testing. Journal of Educational and Behavioral Statistics, 33, 5-20.

Luo Fen, Shuliang Ding & Xiaoqing Wang (2012). Dynamic and Comprehensive Item Selection Strategies for Computerized Adaptive Testing Based on Graded Response Model. Acta Psychologica Sinica, 44, 400-412.

Lu Peng, Dongdai Zhou, Shaochun Zhou & Xiao Cong (2012). Research on Item Selection Method For CAT Based On Simulated Annealing Computer Applications and Software, 29, 175-179.

Brown J.M., Weiss D.J. (1977). An Adaptive Testing Strategy for Achievement Test

Batteries. Research Report 77-6. Minneapolis, Minn: University of Minnesota, Computerized Adaptive Testing Laboratory.

Papoušek, J., Pelánek, R. (2015, June). Impact of adaptive educational system behaviour on student motivation. In International Conference on Artificial Intelligence in Education (pp. 348-357). Springer, Cham.