

# The Design and Implementation of a Computerized Adaptive Test for Student Testing in C++ and Java Exam

Sanja Maravić Čisar\*, Dragica Radosav\*\*, Predrag Pecev\*\*\*\*, Robert Pinter\* and Petar Čisar\*\*\*

\* Subotica Tech/Department of Informatics, Subotica, Serbia

\*\* Technical Faculty "Mihajlo Pupin"/Department of Informatics, Zrenjanin, Serbia

\*\*\* Telekom Srbija, Subotica, Serbia

\*\*\*\* University of Novi Sad, Faculty of Sciences, Dept. of Mathematics and Computers Sciences, Novi Sad, Serbia  
sanjam@vts.su.ac.rs, radosav@zpupin.tf.zr.ac.rs, predrag.pecev@gmail.com, probi@vts.su.ac.rs, petarc@telekom.rs

**Abstract - The use of computerized adaptive testing (CAT) has expanded rapidly over recent years mainly due to the advances in communication and information technology. This paper describes the design issues that were considered for the development of a CAT for student testing programming languages C++ and Java. The application was realized in MATLAB.**

## I. INTRODUCTION

Testing is one of the most common ways of knowledge testing. The main goal of testing is to determine the level of the student's knowledge from one or more subject areas in which knowledge is checked. Different methods of knowledge evaluations are in use, such as in class presentations, writing essays, projects, etc. However, the most common "tool" that is used to test knowledge is the test and oral exam. Since the computer as a teaching tool has been in use more and more in recent decades, and its use has spread to all levels of education, the computer based test has become very popular. This paper presents a computer adaptive test that was realized by the software package Matlab.

## II. THEORETICAL BASIS OF COMPUTERIZED ADAPTIVE TESTS

CAT (Computerized Adaptive Testing) is a type of test developed to increase the efficiency of estimating the examinee's knowledge. This is achieved by adjusting the questions to the examinee based on his previous answers (therefore often referred to as tailored testing) during the test duration. The degree of difficulty of the subsequent question is chosen in a way so that the new question is neither too hard, nor too easy for the examinee. More precisely, a question for which it is estimated, with a probability of 50%, that the examinee would answer correct is chosen. Of course, the first question cannot be selected in this way because at this point nothing is known about the examinee's capabilities (the question of medium difficulty is chosen), but the selection of the second question can be better adapted to each examinee. With every following answered question, the computer is able to evaluate examinee's knowledge increasingly better.

Some benefits of the CAT are [1] as follows:

- Tests are given "on demand" and scores are available immediately.
- Neither answer sheets nor trained test administrators are needed. Test administrator differences are eliminated as a factor in measurement error.
- Tests are individually paced so that an examinee does not have to wait for others to finish before going on to the next section. Self-paced administration also offers extra time for examinees that need it, potentially reducing one source of test anxiety.
- Test security may be increased because hard copy test booklets are never compromised.
- Computerized testing offers a number of options for timing and formatting. Therefore it has the potential to accommodate a wider range of item types.
- Significantly less time is needed to administer CATs than fixed-item tests since fewer items are needed to achieve acceptable accuracy. CATs can reduce testing time by more than 50% while maintaining the same level of reliability. Shorter testing times also reduce fatigue, a factor that can significantly affect an examinee's test results.
- CATs can provide accurate scores over a wide range of abilities while traditional tests are usually most accurate for average examinees.

Despite the above advantages, computer adaptive tests have numerous limitations, and they raise several technical and procedural issues [1]:

- CATs are not applicable for all subjects and skills. Most CATs are based on an item-response theory model, yet item response theory is not applicable to all skills and item types.
- Hardware limitations may restrict the types of items that can be administered by computer. Items involving detailed art work and graphs or

extensive reading passages, for example, may be hard to present.

- CATs require careful item calibration. The item parameters used in a paper and pencil testing may not hold with a computer adaptive test.
- CATs are only manageable if a facility has enough computers for a large number of examinees and the examinees are at least partially computer-literate. This can be a great limitation.
- The test administration procedures are different. This may cause problems for some examinees.
- With each examinee receiving a different set of question, there can be perceived inequities.
- Examinees are not usually permitted to go back and change answers. A clever examinee could intentionally miss initial questions. The CAT program would then assume low ability and select a series of easy questions. The examinee could then go back and change the answers, getting them all right. The result could be 100% correct answers which would result in the examinee's estimated ability being the highest ability level.

The CAT algorithm is usually an iterative process with the following steps:

1. All the items that have not yet been administered are evaluated to determine which will be the best one to administer next given the currently estimated ability level
2. The "best" next item is administered and the examinee responds
3. A new ability estimate is computed based on the responses to all of the administered items.
4. Steps 1 through 3 are repeated until a stopping criterion is met.

Several different methods can be used to compute the statistics needed in each of these three steps, one of them is Item Response Theory (IRT). IRT is a family of mathematical models that describe how people interact with test items [2].

According to the theory of item response, the most important aim of administering a test to an examinee is to place the given candidate on the ability scale [3]. If it is possible to measure the ability for every student who takes, already two targets have been met. On the one hand, evaluation of the candidate happens based on how much underlying ability they have. On the other hand, it is possible to compare examinees for purposes of assigning grades, awarding scholarships, etc.

The test that is implemented to determine the unknown hidden feature will contain  $N$  items, they all measure some aspect of the trait. After taking the test, the person taking the test respond to all  $N$  items, with the scoring happening dichotomously. This will bring a score of either

a 1 or a 0 for each item in the test. Generally this item score of 1 or 0 is called the examinee's item response. Consequently, the list of 1's and 0's for the  $N$  items comprises the examinee's item response vector. The item response vector and the known item parameters are used to calculate an estimate of the examinee's unknown ability parameter.

According to the item response theory, maximum likelihood procedures are applied to make the calculation of the examinee's estimated ability. Similarly to item parameter estimation, the afore-mentioned procedure is iterative in nature. It sets out with some a priori value for the ability of the examinee and the known values of the item parameters. The next step is implementing these values to compute the likelihood of accurate answers to each item for the given person. This is followed by an adjustment to the ability estimate that was obtained which will in turn improve the correspondence between the computed probabilities and the examinee's item response vector. The process is repeated up until it results in an adjustment that is small enough to make the change in the estimated ability negligible. The result is an estimate of the examinee's ability parameter. For each person who is taking the test this process is repeated separately. Nonetheless, it must be pointed out that the basis of this process is that the approach considers each examinee separately. Thus, the basic problem is how the ability of a single examinee can be estimated.

The CAT problems have been addressed before in the literature [3, 4, 5].

### III. THE DESIGN OF COMPUTER ADAPTIVE TEST

For purposes of determining the effects of applying the computer adaptive test for knowledge evaluation, the adaptive test was realized in MATLAB software package (acronym for MATrix LABoratory). MATLAB is an environment for numerical computation and the programming language of MathWorks [6]. MATLAB contains mathematical, statistical, and engineering functions to support all common engineering and science operations. Key features of MATLAB are mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, and numerical integration, 2-D and 3-D graphics functions for visualizing data, tools for building custom graphical user interfaces. Functions for integrating MATLAB based algorithms with external applications and languages, such as C, C++, Fortran, Java, COM, and Microsoft Excel. Add-on toolboxes provide specialized mathematical computing functions for areas including signal processing, optimization, statistics, symbolic math, partial differential equation solving, and curve fitting. MATLAB was chosen because the students during the first year of the study have exercises in this environment in the subject Basic of computing. This subject is mandatory for all students in the first semester. They also use MATLAB during the second and the third year of studying in some specialized subjects.

The computer adaptive test that was implemented in the experiment is a modification of the adaptive test that could be downloaded from the web address [7]. The original test is the adaptive version of the GRE test

(Graduate Record Exam) and runs in a MATLAB command window. The examinee gives his/her answer by typing a letter in front of the answer which they believe is a correct answer. The existing application has been modified to suit the needs of C++ and Java curriculum. The test consists of multiple choice questions with the five possible answers. The appropriate GUI for this application was also realized [8].

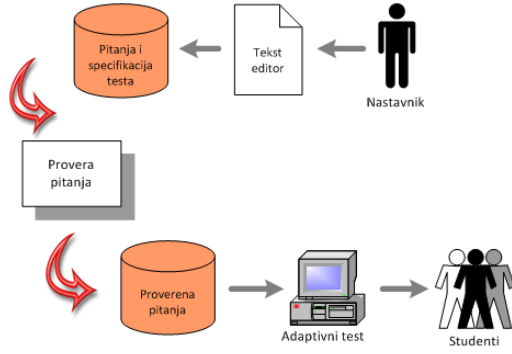


Figure 1. System architecture

The system architecture is shown in Figure 1. The teacher writes questions in some text editor in a pre-define way. For this, only the basic computer literacy is needed and also the basic skill of some text editor (MS Word or Notepad). First, it is necessary to formulate the question, the list of the five possible answers and also the correct answer for that question. The questions for the test are divided into three clusters according to the level of difficulty (easy, medium and difficult), so there are three text files with the questions. All the questions are checked and those ones that passed the control make the question database. From that database questions are given when the test starts. The following section focuses on the explanation of the calibration issues, the assignment of initial question, the choice of the following questions and the rule for stopping the test.

#### A. Item calibration

For each item (i.e. question) in the database it is necessary to perform calibration. According to [9] there are three different approaches: conventional, expert, and online calibration.

The conventional calibration includes methods such as joint maximum likelihood (JML), conditional maximum likelihood (CML), and marginal maximum likelihood (MML). The number of examinees required for the item calibration varies from a minimum 1000 [10, 11], while others recommended between 200 to 1000.

Expert calibration involves calibration of IRT parameters with the use of subject domain experts. The authors in [13] describe the item calibration of 3PL model for computer adaptive test.

Online calibration means using the examinee's response to previously calibrated items is done to estimate parameters of new item during a test.

For creating the database expert calibration was chosen as its implementation was the easiest one. The item calibration is based on Bloom's taxonomy of cognitive skills [14]. During the period of question development and selection for this research, the authors were focused on questions that are capable of giving effective estimation of the first three cognitive skills in Bloom's taxonomy: knowledge, comprehension and application. Table 1 illustrates the question categorization in three clusters according to cognitive skills.

TABLE I. (QUESTION DIFFICULTY ACCORDING TO COGNITIVE SKILLS)

| Težina | Sposobnost    | Kratak opis                                                      |
|--------|---------------|------------------------------------------------------------------|
| 1      | knowledge     | Ability to remember and/or recall previously learnt material     |
| 2      | comprehension | Ability to interpret and/or translate previously learnt material |
| 3      | application   | Ability to apply learnt material in new situations               |

The database consist of questions that were verified in practice (questions from the mid-term tests and the exam in Object-oriented programming and Java) during the period from the year 2005 to 2010, as well as new questions that appeared during the research. The total number of questions was 210 for Object-oriented programming, and 150 for Java. The questions were formulated by the subject's teachers. The questions were classified into three clusters of difficulty as shown in Table 1: 1 (easy), 2 (medium) and 3 (hard). The questions that appeared in the exams in previous years were classified after the statistical processing of the student's answers to them. The new questions, that were created for this research, were classified based on the subjective belief of the authors. The analysis of the experimental results carried out after the research showed that it is necessary to make reclassification of some questions from one cluster to another one. The questions were recorded in .txt files, thus for each cluster there is one .txt file.

#### B. Starting the test

In CAT the question to be selected next for administration depends on the set of previously answered questions. There is a problem of how to select the first question to be administered, although Lord [15] states that, unless the test is very short, the wrong selection of the first question to be administered has little effect on the final result.

One of the possibilities to select the first question was to randomly choose a question from the database. A potential limitation of this choice can be a random choice from either end of the difficulty scale, i.e. choice of very easy or very difficult questions. According to [9] the questions from the end of the difficulty scale are less useful than the questions of medium difficulty. The second approach is to select a question based on

information about the examinee, such as scores on previous tests, or scores in similar subjects. Because this historical information about the examinee is not always available, it was chosen to start the test with the question of medium difficulty (cluster 2).

#### C. *Selecting the next item to be administered*

The test starts with a question of medium difficulty. If the examinee answers this question correctly, the next question to be administered is the first question from cluster 3, i.e. the level of difficulty is one level higher. If the examinee answers incorrect on the first question, the next question is the first question from the cluster 1. After each given is checked and based on that the next question is selected to be administered. Also, after each answer it is checked if the total amount of questions (in this case 20) is reached. If the student gave an answer to all 20 questions, the final score is calculated and shown to the student.

#### D. *Zaustavljanje testa Stopping the test*

The CAT can be completed when answers are given to a fixed, predetermined number of questions, or if the pre-scheduled time for the test has expired. This kind of test is called computer adaptive test of fixed length. Another possibility is for the adaptive test to be considered complete when it reaches a satisfactory level of measurement's precision, for example when the standard error of estimating the ability of the examinee reaches a pre-defined level. This kind of test is called computer adaptive test of variable length. It is also possible to define a stopping rule that is a combination of the two rules, for example, the test is completed when the examinee answered all questions provided for administration, or when time runs out to for solving a specific test, depending on which of these two conditions are met first.

In [11] the authors pointed out that if the length of the test is variable, the examinees with low abilities have to answer much shorter tests than those who are skilled, while in [16] the authors indicate that examinees have doubts about the tests that are very short. The different length of the test raises doubts among the examinees regarding the fairness of grading.

Computer adaptive tests of fixed length are easier to implement, and also if the number of questions is fixed it is much easier for the examiner to predict how many questions need to be in the database. In [17] the authors recommended three to four times more questions in the database than the number of questions required for the test.

Having in the mind all the previously stated, the test with fixed length of 20 questions was selected with limitations of 30 minutes for completing the test.

The interface of the application is described in [18].

#### IV. IMPLEMENTATION

The research was done at Subotica Tech – College of Applied Sciences, Subotica, Serbia. The total number of students who participated in the research was 352, representing 48.35% of the total number of students at Subotica Tech.

The results of the research are shown with a certainty of 95% and the risk of 5%, namely, that in six of eight cases examined there is a statistically significant difference between the results of the control group (the group working on the conventional test method) and the experimental group (adaptive test), i.e. students who have done a computer adaptive test achieved a better result compared to students who have done the test in a conventional way. The average difference in the test score was about 10 points in favor of the students of the experimental group.

Based on student responses to a questionnaire regarding their opinion about the computer adaptive test, it is clear that there is a positive attitude towards it. Asked whether they would recommend a computer adaptive test to colleagues, 86.90% students of experimental group had affirmatively attitude, while an equal number of students were undecided or would not recommend it. More than 60% of students found that this method of taking the tests is less stressful than the classical paper-and-pencil test. Based on the comments that were written by students it may be stated that this way of taking the test is more comfortable because it adapts to their knowledge and skills, thus reducing the frustration when the questions are too difficult for some students, or boredom that occurs with good students when questions are not challenging (too easy) for them. In both cases, there could be the problem of demotivation and dissatisfaction, as well as the lack of desire for progress.

#### V. CONCLUSION

During the process of solving the classical test, students may feel discouraged if the questions are too difficult, or, on the other hand, they may lose interest if the questions are too easy for their level of knowledge. The solution to this problem may be the application of computer adaptive tests (CAT), which, along with quality oral assessment, have the option to alter the level of questions difficulty to the capabilities of the respondents.

The research results provide the basis for further work which would be directed towards improving the application by adding multimedia elements (images, sound and video). Also, the improvement of the application should go into the direction of realization of the feedback to the student. The function of feedback is not only to give indications that the answer is true or not, but also to point students towards the lesson in which there is an answer to the question at hand.

#### REFERENCES

- [1] <http://echo.edres.org:8080/scripts/cat/catdemo.htm>
- [2] S. E Embretson, and S. P. Reise, "Item response theory for psychologists", Mahwah NJ, Lawrence Erlbaum Associates, 2000.
- [3] S. Kardan, and A. Kardan, "Towards a More Accurate Knowledge Level Estimation", 2009 Sixth International Conference on Information Technology: New Generations, Las Vegas, Nevada, pp. 1134-1139, 2009, <http://doi.ieeecomputersociety.org/10.1109/ITNG.2009.154>
- [4] Gin-Fon N. Ju, A. Bork, "The Implementation of an Adaptive Test on the Computer," icalt, Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05), Kaohsiung, Taiwan, pp. 822-823, 2005, <http://doi.ieeecomputersociety.org/10.1109/ICALT.2005.274>

- [5] F. Baker, "The Basics of Item Response Theory", chapter 5, Estimating an Examinee's Ability, <http://echo.edres.org:8080/irt/baker/chapter5.pdf>, 2001.
- [6] <http://www.mathworks.com>
- [7] Mathworks, Computer Adaptive Test Demystified, <http://www.mathworks.com/matlabcentral/fileexchange/12467-computer-adaptive-test-demystified-gre-pattern>
- [8] Jekić, R. (2010). GUI za kompjuterizovano adaptivno testiranje izrađen u Matlabu, seminarski rad, VTS.
- [9] Lilley, M. (2007). The Development and Application of Computer Adaptive Testing in a Higher Education Environment, PhD thesis. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.100.7459&rep=rep1&type=pdf>
- [10] Wainer, H. & Mislevy, R. J. (2000) Item Response Theory, Item Calibration, and Proficiency Estimation, in H. Wainer (2000), Computerized Adaptive Testing: A Primer, Lawrence Erlbaum Associates Inc.
- [11] McBride, J. R. (2001) Technical Perspective, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), Computerized adaptive testing: From inquiry to operation, Washington DC: American Psychological Association.
- [12] Huang, S. X. (1996). A Content-Balanced Adaptive Testing Algorithm for Computer-Based Training Systems, Lecture Notes in Computer Science, 1086, pp. 306-314.
- [13] Conejo, R.; Millán, E.; Pérez-de-la-Cruz, J. L.; Trella, M. (2000) An Empirical Approach to On-Line Learning in SIETTE, Lecture Notes in Computer Science, 1839, pp. 605-614.
- [14] Bloom, B.S. (Ed.) (1956). Taxonomy of educational objectives: The classification of educational goals: Handbook I, Cognitive domain. New York;Toronto: Longmans, Green.
- [15] Lord, F. M. (1980). Applications of item response theory to practical problems. Hillsdale, NJ: Lawrence Erlbaum Associates.
- [16] Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage Press.
- [17] Carlson, R. D. (1994). Computer-Adaptive Testing: a Shift in the Evaluation Paradigm, Journal of Educational Technology Systems, 22(3), pp 213-224.
- [18] Maravić Čisar, S., Radosav, D., Markoski, B., Pinter, R., Čisar, P. (2010). Computer Adaptive Testing of Student Knowledge, Acta Polytechnica Hungarica, Vol.7, No.4, ISSN 1785-8860, pp. 139-152.