Abstract—Computer based tests are becoming increasingly important and are thus implemented in all fields of education. The clear advantage of computer adaptive testing is witnessed by the maximal balance of accuracy and efficiency. For purposes of determining the effects of applying the computer adaptive test for knowledge evaluation, the adaptive test was realized in the MATLAB software package. The research was done at Subotica Tech – College of Applied Sciences (Subotica, Serbia) with students from the Department of Informatics. The MATLAB application was compared with two computer adaptive tests with a similar purpose: testing student’s knowledge of programming languages. The comparison was done for the four basic components that every computer adaptive test has to have: item bank, starting point, item selection algorithm and stopping rule.

I. INTRODUCTION

One of the most common ways of determining the measure of knowledge acquisition in a certain field of science is by testing. A great number of methods are used in areas as various as class presentations, writing essays, projects, etc. The test and the oral exam are the commonest used in knowledge testing. Computer based tests are becoming increasingly important and are thus implemented in all fields of education. The clear advantage of computer adaptive testing is witnessed by the maximal balance of accuracy and efficiency. The CAT (Computer adaptive test) has been applied more and more in the last quarter century covering the fields of education, certification, and licensure [1].

For purposes of determining the effects of applying the computer adaptive test for knowledge evaluation, the adaptive test was realized in the MATLAB software package. The application was done in Matlab based on the program code that can be found under the web address in [2]. The original code presents a computer adaptive test for GRE (Graduate Record Exam) and enables questions of the following types: analogy, antonym, and fill in the blanks. It was modified to allow testing basic concepts of C++ and Java in the form of multiple choice questions. The research was done at Subotica Tech – College of Applied Sciences (Subotica, Serbia) with students from the Department of Informatics. Some results of the research, as well as description of the application can be found in previous publications [3, 4, 5]. This application was compared with two computer adaptive tests with a similar purpose: testing student’s knowledge of programming languages. The comparison was done for the four basic components that every computer adaptive test has to have: item bank, starting point, item selection algorithm and stopping rule [6].

The remainder of this paper is organized as follows: Section 2 briefly reviews the four basic components of CAT, while in Section 3 the brief description of selected computer adaptive tests are presented, then the conclusion is given at the end.

II. THE BASIC COMPONENTS OF CAT

Every CAT is composed of five components. The first component is the calibrated item bank, then starting point, item selection algorithm, scoring algorithm and criterion for termination the test. The scoring algorithm was not taken into account in the comparison of the tests.

A. Item calibration

Calibration has to be performed for each item (i.e. question) in the database. Reference [7] classifies three different approaches: conventional, expert, and online calibration.

Joint maximum likelihood (JML), conditional maximum likelihood (CML), and marginal maximum likelihood (MML) are among the conventional calibration methods. There are different views concerning the number of examinees required for the item calibration, it ranges from a recommended 200 to 1000 to a minimum of 1000 [8, 10].

In order to achieve expert calibration one has to calibrate of IRT parameters with the use of subject domain experts. Reference [11] presents a description of the item calibration of 3PL model for computer adaptive test.

In order to estimate parameters of a new item during a test one uses the responses to previously calibrated items as a form of online calibration.

B. Starting the test

The set of the previously answered questions are taken as basis for the next question for administration to be selected in CAT. However, this leaves the challenge of how to make the selection of the first question. In [12] Lord proposes that even if the wrong decision is made regarding the first question, it will have little impact on the eventual result, unless the test is not long.

The options for choosing the first question included the random selection from the database. However, this includes the risk of selecting the first question from either end of the difficulty scale, i.e. so it could either be a very easy or very difficult question. Reference [7] argues that medium difficult questions are more useful than those from the other end of the difficulty scale. One can also take a different approach, namely to determine the first question using some information discovered about the examinee, including how much they scored on previous tests or in similar subjects.
C. Stopping the test

The condition for completing the CAT is for a fixed, predetermined number of questions to be given or to remain within the pre-determined time schedule. This sort of test is defined as a fixed-length computer adaptive test. On the other hand, a test is also seen as complete when an acceptable level of measurement precision is reached, when the standard error of estimating the ability of the examinee reaches a pre-defined level. This is how the variable length computer adaptive test is defined. A stopping rule can also be implemented, which is the combination of the two rules, such as deeming the test as complete when either all questions are answered as provided for administration, or the given time has expired, depending on which of these events occurs first.

Authors in [10] found that with variable length tests, low-ability examinees are given much shorter tests than skilled examinees. In reference [13] authors argue that very short tests are met with doubt, varying length tests raised questions regarding the fairness of grading.

In terms of implementation, fixed-length tests are easier, as it makes it easier for the examiner to predict how many questions are needed in the database. According to [14] the number of questions recommended for the database is three to four times the amount of questions required for the test.

III. DESCRIPTION OF CAT

In this section three CATs are described and compared: PAT, MATLAB and an application from the University of Hertfordshire.

A. Programming Adaptive Testing –PAT (Greece)

Programming Adaptive Testing (PAT) is a system for adaptive testing that was developed at University of Macedonia, Information Systems Department, Thessaloniki, Greece. The test is for beginners in programming. The Flash MX and the Action Script was used for developing this computer adaptive test. The application is design for testing knowledge of programming and for classifying students based on their programming skills in the three categories: beginners, medium and advanced group of students. The aim of testing was to predict the result of students on Panhellenic National Exam based on testing the programming skills [15].

The questions were divided into three levels of difficulty: A=easy questions, B=medium, and C=difficult questions. The questions were designed according to the first three cognitive domains of Bloom’s taxonomy: knowledge, comprehension and application. For the first cognitive domain, which is related to recalling data or information, the questions were about the syntax of programming language Glossa that students learned, fundamentals of structured programming and subroutines.

The questions for the second cognitive domain were about understanding the part of the programming code. For example, finding out the output after execution of the programming code. These types of questions were for the B and C levels of difficulty.

The questions that were designed for the application on what was learned in the classroom, demand selections of the appropriate equivalent command for some part of the programming code, selection of commands that is represented by a logical diagram, as well as filling up gaps in the programming code. These types of the questions belong to the C level. The total number of items in the database amounts to 443.

There were two types of the questions: the multiple choice questions (MCQ) and the true/false questions. The true/false questions were not intended for use for evaluation of knowledge from the third Bloom’s cognitive domain. The MCQ differ in the number of possible answers depending on which Bloom’s cognitive domain they were design. The MCQ for the level A have 3 possible answers, for the level B four and for the level C five possible answers.

The test starts with a random selection of a question from the A level. If the examinee answers this question correctly, the next question to be administered is the question from the next level of difficulty. If the examinee answers incorrectly, the next question to be administered is a question from the lower level of difficulty. Within each level of the difficulty, the selection of the next question is random. The total number of the questions on the test is 30. There is no possibility to return to the previous question, which is one of the general characteristics of computer adaptive tests. The example of PAT test is shown in Fig.1.

![Image 1](Image 1.png)

The number of points for the each question depends on the level of difficulty in which the question belongs. The correct answer to the question from the A level is worth 1 point, from the B level 2 and for the C level 3 points [15]. After finishing the test, the examinee gets a report about his/her success on the test as shown in Fig. 2.

![Image 2](Image 2.png)
The examinee receives the following information:

Total score (x): the number of correct answers from the total number of 30 questions. The maximum value is 30.

The number of correct answers are presented per level, given by the total number of questions asked at a given level. For example, 6/24 means that the examinee had 6 correct answers of 24 questions that were asked at a certain level.

The final result (y) is calculated as follows: the number of correct answers at the level of A + the number of correct answers at the level of B + the number of correct answers at the level of C. The maximal value is 87.

The classification of examinees according to the programming skills in one of three categories is done as follows:

- If \(0 \leq x \leq 17\) and \(0 \leq y \leq 33\) a low level of programming skills (beginners)
- If \(16 \leq x \leq 20\) and \(34 \leq y \leq 51\) a medium level of programming skills (intermediate)
- If \(21 \leq x \leq 30\) and \(52 \leq y \leq 87\) a high level of programming skills (advanced)

B. CAT from the University of Hertfordshire (Great Britain)

The mathematical model that was used for this application is based on the three-parameter logistic model (3PL) of the item response theory (IRT). In the 3PL model, the probability of a correct response to an item is:

\[
P(\theta) = c + \frac{1-c}{1+e^{-\frac{(\theta-b)}{a}}}
\]  

where \(\theta\) is the person (ability) parameter and \(a\) (the discrimination parameter), \(b\) (the difficulty parameter) and \(c\) (the guessing parameter) are the item parameters [16].

For each examinee’s answer, a mathematical function for 3PL model is used for estimating the level of ability of the examinee. It is necessary to store information about \(a\), \(b\) and \(c\) parameter for each question in the item database [11]. For the questions without historical data, the initial value of the difficulty parameter \(b\), was define based on the expert’s estimation from -3 (minimum value) to +3 (maximal value). The item calibration was done based on Bloom’s taxonomy [19] and is given in Table 1.

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Cognitive skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3 ≤ b ≤ -1</td>
<td>knowledge</td>
</tr>
<tr>
<td>-1 ≤ b ≤ +1</td>
<td>comprehension</td>
</tr>
<tr>
<td>+1 ≤ b ≤ +3</td>
<td>application</td>
</tr>
</tbody>
</table>

The test was first divided by chapters, and then for each chapter, questions were defined for every of three previously defined levels of difficulty. The total number of items in the database was 150. The test starts with a question randomly selected from the group of questions with \(b=0\) (medium level of difficulty). Based on the student’s answer, the level of ability is estimated according to the (1) and then a question which has the parameter \(b\) closest to the estimated level of ability is selected from the database. The stopping rule is a combination of the two rules: the test is completed when the examinee answered all questions provided for administration, or when the time for solving a specific test runs out, depending on which of these two conditions is met first. The application was done using Visual Basic 6 for Microsoft Windows [9].

The questions were MCQ with the four possible answers. A screenshot of the test is shown in Fig.2.

After completing the test, examinee receive a report about the success on the test, as shown in Fig.3.

<table>
<thead>
<tr>
<th>Your Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name:</td>
</tr>
<tr>
<td>Your score: 75%</td>
</tr>
<tr>
<td>Module:</td>
</tr>
<tr>
<td>Assignment: Human-Computer Interaction (HCI) Multiple-Choice Test</td>
</tr>
<tr>
<td>Weighting (out of final mark): 10%</td>
</tr>
</tbody>
</table>

The answers are grouped according to the chapters where the question belongs to, after that the level of ability is estimated for each chapter, as shown in Fig.4. The estimated ability then is mapping according to Bloom’s taxonomy of cognitive skills. The examinee is given information about his/her success on the test (the red arrow) compared to the average success of the group (the blue rectangle).

The final score according to the Bloom’s taxonomy [9]
C. MATLAB CAT (Serbia)

This current experiment implemented a computer adaptive test that was somewhat modified from the CAT found under the web address [2]. The original one is adapted from the GRE test (Graduate Record Exam) and can be run in a MATLAB command window. The answers are given by the examinee entering a letter before the answer assumed to be correct. Some modifications were carried out in order to accommodate the needs of C++ and Java curriculum. There are multiple choice questions in the test, each with the five possible answers. The application also includes the appropriate GUI.

The questions are typed into a text editor by the teacher, which requires nothing more than basic computer literacy as well as some text editor knowledge (MS Word or Notepad). The first step is to formulate the question, then provide five possible answers as well as the correct answer for the given question. All questions are classified into three groups based on their level of difficulty, ranging between easy, medium and difficult. This results in three separate text files with the questions. When the test starts, the questions are drawn from that database. Now follows the description of the calibration issues, how the first question is assigned, how the subsequent question is selected and what the rules for stopping the test are.

Since it was the easiest for implementation, expert calibration was chosen for creating the database. Bloom’s taxonomy of cognitive skills was used as a basis for item calibration [19]. In the phase of developing and selecting the questions for this experiment the spotlight was on those questions which could efficiently estimate the first three cognitive skills in Bloom’s taxonomy: knowledge, comprehension and application. How the questions were categorized into three groups based on cognitive skills can be seen in Table 2.

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Ability</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>knowledge</td>
<td>Ability to remember and/or recall previously learnt material</td>
</tr>
<tr>
<td>2</td>
<td>comprehension</td>
<td>Ability to interpret and/or translate previously learnt material</td>
</tr>
<tr>
<td>3</td>
<td>application</td>
<td>Ability to apply learnt material in new situations</td>
</tr>
</tbody>
</table>

Question would only make the cut into the database once they were verified in verified in practice, thus they would be questions previously part of the mid-term tests and exams in the subjects Object-oriented programming and Java, between 2005 and 2010. The database also included new questions. The questions were taken from two subjects, Object-oriented programming with 210 and Java with altogether 150 questions. The teachers of these subjects were responsible for wording the questions. There are three clusters of difficulty which the questions are grouped in, presented in Table 1: cluster 1 with the easy questions, cluster 2 with the medium, and 3 with the hard questions. Questions from the previous years’ exams were categorized based on the statistical processing of the student’s answers to them. Question storage was in the form of .txt files resulting in one .txt file for each cluster.

The initial question of the test is one of medium difficulty (cluster 2). Supposing the student’s answer is correct, the subsequent question will be of greater difficulty, so from cluster 3, one level higher. The implemented algorithm ensures that after two consecutive incorrect answers of the same level of difficulty the next selected question comes from the cluster of one level of difficulty lower. So, in the best case, when the examinee answers all questions correctly, the scenario would be as follows: 2T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T, 3T.

An overview of the four components that every computer adaptive test has to have (item bank, starting point, item selection algorithm and stopping rule) for PAT, MATLAB CAT and adaptive test from the University of Hertfordshire is given in Table 3.

![Figure 6. The MATLAB CAT](image-url)
If these applications are compared with the computer adaptive test realized in MATLAB, one finds that they are all designed for beginners in programming. According to that, questions were selected to give estimations of the first three cognitive levels of Bloom’s taxonomy: knowledge, comprehension and application. The questions were categorized into the three groups of difficulty: easy, medium and advanced. The type of questions that is used is multiple choice questions (MCQ). In the PAT application, the number of possible answers depends on the category which the question belongs to (3, 4 or 5), in the Hertfordshire application there are four possible answers and in MATLAB, there are five.

The test starts with a question of medium level of difficulty in Hertfordshire and MATLAB application, while in the PAT it is a question from the group of the easiest questions. The number of questions in the database in all three applications satisfies the condition that the number of questions should be 5 to 10 times larger than the number of questions in the test.

PAT and MATLAB are adaptive tests of fixed length (given answer to all questions) and the application from the University Hertfordshire uses a stopping rule that is a combination of the two rules (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).

The biggest difference between these applications is the algorithm that is applied to select the next question. In PAT application, if the answer is correct, the level of difficulty for the next question is one level higher and in application from the University of Hertfordshire 3PL model and IRT is applied. In MATLAB application, after two consecutive incorrect answers of the same level of difficulty, select as a next question, question from the cluster of one level lower. If the answer is correct, the level of difficulty is one level higher.

### IV. Conclusions

For the purposes of determining the effects of applying the computer adaptive test for knowledge evaluation, the adaptive test was realized in MATLAB software package. MATLAB was chosen because during the first year of their studies the students 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 authors believed that the working environment that is familiar to all students would contribute to a positive attitude and reduce the tension during the test.

The authors are aware of the shortcomings of the applied solutions in terms of HCI, but the priority in the development of applications was the speed of learning. The number of students that participated in the research was more than 300 [21] and it would take a long time on their training, the authors have opted for a minimalist design in a familiar environment to all students (MATLAB). The aim was: achieving a low level of errors by the users, offering a user-friendly interface without additional icons and buttons so that users can focus on the question. In the later stages of application development, one must take into account the development of applications according to HCI principles.

### REFERENCES


computer-adaptive-test-demystified-gre-pattern


