Computer Adaptive Testing for Student's Knowledge in C++ Exam

Sanja Maravić Čisar*, Dragica Radosav**, Branko Markoski**, 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
sanjam@vts.su.ac.rs, radosav@zpupin.tf.zr.ac.rs, markoni@uns.ac.rs, probi@vts.su.ac.rs, petarc@telekom.rs

***Abstract*—In adaptive learning systems for distance learning attention is focused on adjusting the learning material to the needs of the individual. Adaptive tests adjust to the current level of knowledge of the examinee and is specific for their needs, thus it is much better at evaluating the knowledge of each individual. The basic goal of adaptive computer tests is to ensure the examinee questions that are challenging enough for them but not too difficult, which would lead to frustration and confusion. The aim of this paper is to present a computer adaptive test (CAT) 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

978-1-4244-9280-0/10/$26.00 ©2010 IEEE

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 estimation equation used is shown below:

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \frac{\sum_{i=1}^{N} -a_i[u_i - P_i(\hat{\theta}_s)]}{\sum_{i=1}^{N} a_i^2 P_i(\hat{\theta}_s) Q_i(\hat{\theta}_s)} \quad (1)$$

where: $\hat{\theta}_s$ is the estimated ability of the examinee within iteration s, a_i is the discrimination parameter of item *i*, *i = 1, 2,... .N.*

u_i is the response made by the examinee to item *i*:

$u_i = 1$ for a correct response

$u_i = 0$ for an incorrect response

$P_i(\hat{\theta}_s)$ is the probability of correct response to item *i*, under the given item characteristic curve model, at ability level $\hat{\theta}$ within iteration *s*.

$Q_i(\hat{\theta}_s)=1 - P_i(\hat{\theta}_s)$ is the probability of incorrect response to item *i*, under the given item characteristic curve model, at ability level $\hat{\theta}$ within iteration *s*.

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

III. DESCRIPTION OF THE APPLICATION

The application was done in Matlab based on the program code that can be found at the web address [6]. The program presents a computer adaptive test. For demonstration purpose, GRE (Graduate Record Exam) is taken. The program was modified to enable the testing of student knowledge in C++. The application can run from a Matlab command window, or it is possible to make a stand alone application that does not require the installation of Matlab.

After starting the program the main window is displayed and the dialog for entering basic data on student (name, surname, index number). Pressing the Enter command button starts the test, as shown in Figure 1.

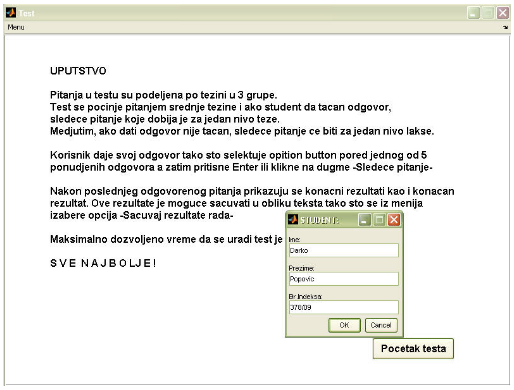

Figure 1. Startup screen

After registering for the test a new window opens with the first question. At all times the student can see on the screen which question the student is on, the total number of questions, the text of the question with multiple choice answers, as can be seen in Figure 2. At the bottom of the screen there is a progress bar which illustrates the progress of the student during the test.

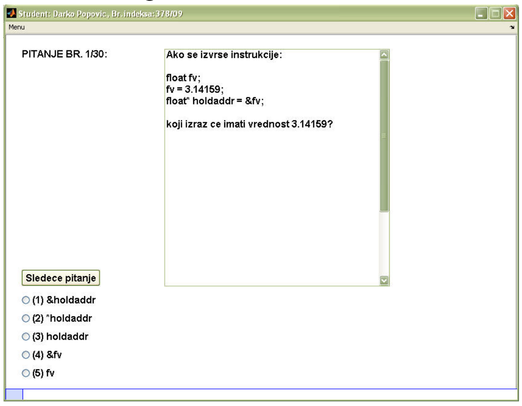

Figure 2. Screenshot of a question

The questions are divided based on their difficulty into three groups, easy, medium and difficult question. The test starts with a question of medium difficulty, which is assigned in test the results with number 2. If the student gives a correct answer to this question, the algorithm of the test passes to the first question in the group of difficult questions (assigned with number 3), and if the answer given to the first question is incorrect then the group with easy questions is selected for (assigned with number 1). The questions are written in a txt file and they are invoking by calling the appropriate function in the program. This allows the examiner to easily modify and extend the question database, without knowledge of the syntax of any programming language, but it is only required to follow a determined format of questions writing.

After answering the last question, the examinee can see their results immediately on the screen. If the examinee selects the option to save the test results, the appropriate function parameters would be call. From the text file the level of the question's difficulty can be seen, whether or not the answer was correct or incorrect, and the time needed for answering each question (i.e. until pressing the command button Next question/Show results).

	Tezina pitanja	Tacno/Netacno	sec/pitanju
1	2	0	19.9536
2	1	0	6.3628
3	1	0	9.1372
4	1	1	5.3806
5	2	0	5.9608
6	1	0	3.9407
7	1	1	7.4421
8	2	1	5.5676
9	3	0	6.0992
10	2	1	5.8967
11	3	1	19.6430
12	3	1	8.3513
13	3	1	3.3496
14	3	1	5.3362
15	3	0	21.7765
16	2	0	7.5058
17	1	0	5.6952
18	1	1	8.2771
19	2	0	8.7138
20	1	1	15.5513
21	2	0	7.1675
22	1	1	9.3100
23	2	1	4.1316
24	3	0	7.8651
25	2	1	10.2798
26	3	1	6.1906
27	3	1	6.9241
28	3	0	6.5549
29	2	1	5.3610
30	3	0	11.0152

Figure 3. Test results view

IV. CONCLUSIONS

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.

This paper reports on the use of a computer adaptive test for student's knowledge testing in C++. The motivation behind this work was to investigate techniques for the improvement of student assessment. Future work will involve the further analysis of the test statistics and improvement of the classification of questions based on the student's test results.

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] Mathworks, Computer Adaptive Test Demystified, http://www.mathworks.com/matlabcentral/fileexchange/12467-computer-adaptive-test-demystified-gre-pattern