MEST JOURNAL
DOI 10.12709/mest.01.01.02.02

KNOWLEDGE ASSESSMENT AND APPLICATION OF COMPUTER ADAPTIVE TESTING

Zoran Čekerevac
Faculty of Business and Industrial Management, "Union" University Belgrade, Belgrade, Serbia

Svetlana Andjelić
Information Technology School - ITS, Beograd, Serbia

Petar Čekerevac
Faculty of Political Sciences, University of Belgrade, Belgrade, Serbia

JEL category: **D8, D81, C1, C12**

Abstract

Assessment of knowledge and skills of students, and people at all, are everyday practice in today's world. It is implemented in different ways and for that purpose many methods are used. It is impossible to find one general method, the best in all circumstances. The first part of this paper deals with the problems that arise in the evaluation, and the second part analyzes the possibility of using of computer adaptive testing to determine the grades of students' knowledge. Special attention is paid to the objectivity of evaluation and to the impact of assessors on the final mark. Here are reconsidered examples of grading of the students' papers presented and defended in an international student's competition, as well as defending of several students' term papers in the frame of regular university classes. The results were compared, and it was pointed to many of aspects of establishing the objectivity of the obtained scores. In the section dealing with the presentation of the application of computer adaptive testing in the evaluation, there are presented algorithm of testing, research methods and results of the research on concrete examples from practice. Here is, also, made the comparison of the scores that the same students had achieved in two testing modes: classical testing and computerized adaptive testing. In addition to this comparison, this paper presents the results of a survey of the impressions of tested students. In doing so, the questions were focused on the appropriateness of computer adaptive testing and the other impressions that students had gained during testing, satisfaction with the earned mark and so on. The results achieved by applying computer adaptive testing are summarized in the conclusions, and also some advantages and disadvantages of

The address of the coresponding author:
Zoran Čekerevac
zoran@cekerevac.eu

such testing are discussed. At the end, the paper answers the question about the usefulness of evaluations and, also, some questions about the objectivity of the assessment.

Keywords: Knowledge assessment, CAT, computer adaptive testing, estimation of knowledge, Bayes' theorem, MAP approach, spearman rank-order correlation coefficient, grading system

1 INTRODUCTION

Assessment of knowledge and skills has always been an important activity in teachers' and students' life and work. Even not enrolled in a college, prospective students meet the challenges of exam which should show whether the high school graduates are able to attend classes at the university. Based on the presented results the lists are formed, and they are the basis for the selection of candidates. Since universities are aware that the exams are not sufficiently reliable criterion for measuring of the quality of incoming students, they introduce the other criteria which should throw more light on the abilities of candidates, for example, success in previous high school. However, when all factors are analyzed, it is not difficult to conclude that a major influential factor for admission to a college is the result on the entrance exam.

When a candidate becomes a student, during the study, he has to expect a series of examinations that should indicate to what extent the student mastered the skills in certain fields.

And here, universities are aware that only the test passing need not to be a real measure of student knowledge and skills, and they, mostly for themselves, are looking for ways to objectify grades. This independence has created a great diversity in assessments, and also a lot of difficulties during the comparison of the results achieved of graduates. At the contests the reputation of the university is usually taken as the first criterion, and then, as the second criterion, the results of students' work. It is therefore important to find an answer to questions of expediency and accuracy of assessment and evaluation personalization.

In the last fifteen years, primarily due to the formation of the European Union, in order to equalize the quality of study and assessments, the EU states acceded to the change of the studying system. The results of these activities are the Joint declaration of the European Ministers of Education convened in Bologna on the 19th of June 1999 (EU, The Bologna Declaration, 1999), as the main guiding document of the Bologna process, and the Prague communiqué from the year (2001), the Berlin communiqué (2003), the Bergen communiqué (Bergen, 2005), the London communiqué (2007) the Leuven & Louvain-la-Neuve communiqué (2009), as well as the Budapest-Vienna Declaration (2010) and many others. Many additional criteria were introduced in order to improve the quality of studying and the results achieved. However, the examination remained the most important measure of achievements and acquired knowledge and skills.

Considering that the final score is authoritative, on the basis of the obtained scores students gain their starting positions in their future work which undoubtedly affects their future development and their future life at all. Students who have received better grades for their work can also expect better jobs at the beginning of their careers. However, one can always ask the question: How real and how exact is the grade of gained knowledge? During the examination, student answers to the several questions that cover some parts from the studied subject's program. Has the student mastered the complete subject matter on the same level, and it does not matter on which question student gives his answers, or maybe, student has mastered some parts better than others? Has the student got questions from the matter he mastered better, or, maybe, from the matter he mastered poorly? That is not known to anyone, except maybe, to the student, who possibly, in case that he thinks that the grade he got is low, may cancel the exam.

2 THE OBJECTIVITY OF MARKS IN THE CLASSICAL EVALUATION

While analyzing student marks that have been given on the classical exams, some hypotheses can be set:

Working hypothesis #1

H_0 The score that the student gets on the exam depends on examiner.

Alternative hypothesis #1

H_1 All assessors under the same conditions for the same work give the same marks.

In the case of assessing of students by use of written exams, arithmetic problems or test, all students receive the same questions, so the student may answer to the question well, wrong, or to partially good. Based on the given responses, teacher forms the final mark that defines the level of achieved student knowledge. Taking into account that, for checking of the null hypothesis #1, non-parametric correlation should be considered, the conclusion imposes, that for verification of the hypothesis it is necessary to calculate Spearman Rank-Order Correlation Coefficient. To calculate this coefficient it is not necessary that the data are normally distributed and linear correlated, while the sample size may be smaller than 35. (Dawson & Trapp, 2004).

In classical testing, as a material for the assessment, the teacher uses a written document. Also, the teacher himself may affect to the final mark. The teacher, as well as any other person, can be affected by factors such as fatigue, mood - bad mood, previous experience with the student and so on.

In addition to that, the teacher alone may have a different relationship to the different parts of the matter that he taught. He can consider some parts as significant, and some parts as less important. Issues from some parts of the subject can be more interesting to him, and asked about more frequently, while some parts of the curriculum he can consider less important, and, consequently, does not ask these questions. If students have such experiences with the teacher, they can avoid certain parts of the subject matter, and give greater attention to other parts of the curriculum. Finally, different teachers have different criteria, so it can be assumed, that, for the same level of student's knowledge, the marks may be different by different examiners.

As an illustration may serve the results of a students' competition, held in Bratislava on VSEMVS in spring 2012. The competition was attended by 13 students of master studies, and their papers were assessed by an international jury of five renowned professors. The objective of the assessment was to rank papers and their presentations by the quality, and to award the five best papers. The results of evaluation are shown in the Table 1.

Table 1 The results of evaluation

	Paper	Grader #1	Grader #2	Grader #3	Grader #4	Grader #5	Sum	Final positions of papers
1	Paper #01	35	53	52	41	38	219	4
2	Paper #02	33	38	55	44	33	203	7
3	Paper #03	40	52	51	40	34	217	5
4	Paper #04	41	41	49	48	34	213	6
5	Paper #05	41	41	40	35	32	189	10
6	Paper #06	38	45	55	53	36	227	3
7	Paper #07	30	38	30	31	32	161	13
8	Paper #08	33	40	46	47	34	200	8
9	Paper #09	56	57	60	56	37	266	1
10	Paper #10	34	52	56	43	35	220	2
11	Paper #11	34	52	35	32	38	191	9
12	Paper #12	18	51	35	32	36	172	11
13	Paper #13	42	44	40	37	37	200	8
14	Paper #14	27	39	29	35	32	162	12

Table 1 shows the grades that five assessors gave, in points, for each paper. The scores are then added and the totals for each of the papers were ranked in the column „**Final positions of papers**".

Looking at the ratings by assessors, it can be observed that the average score of all the papers of different assessors is different and ranges from 34.9 (the assessor #5) to 45.9 assessor #2). Figure 1 shows average marks of all papers by assessors.

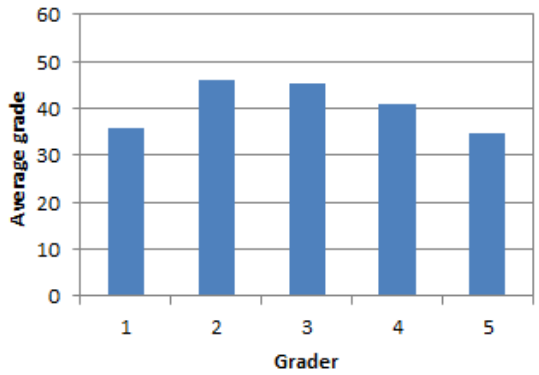

Fig. 1 Average grade per grader

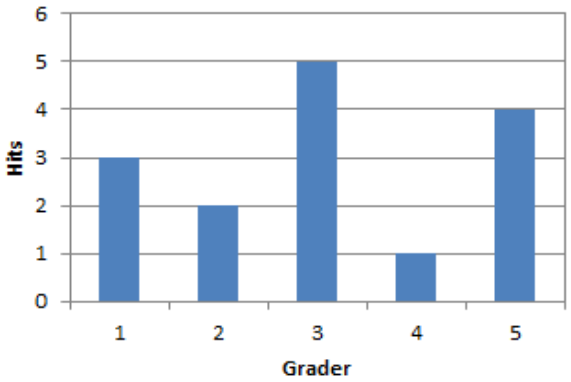

Fig. 2 Number of grader's hits

Comparing rankings of the papers based on evaluations of individual graders with the order of papers in the final ranking, one can observe that the graders' rankings differ considerably from the final ranking. Thus, the first grader's ranking in three positions coincides with the final ranking. Rankings of other assessors are consistent with the final ranking list in 2, 5, 1, 4 positions, as shown on Figure 2. In addition, different assessors gave to the same work and to the same oral presentation a very different numbers of points. It is possible to see for the five best ranked papers in Figure 3, and for all papers in Table 2. Mean interval of variation for all fourteen papers was 17.14, which would correspond, in 10-grades rating system, to the span of two grades. The maximal interval of variation was 33 points, at paper #12, which would correspond, at the same system, to the range of more than three grades. Even greater differences arise when comparing intervals of variations with the average grades of papers. The smallest deviation is in the case of the paper #13 in the amount of 17.50%, and highest, in the case of paper #12, that amounts 95.93%.

Based on the sample used there is no reason to reject the null hypothesis.

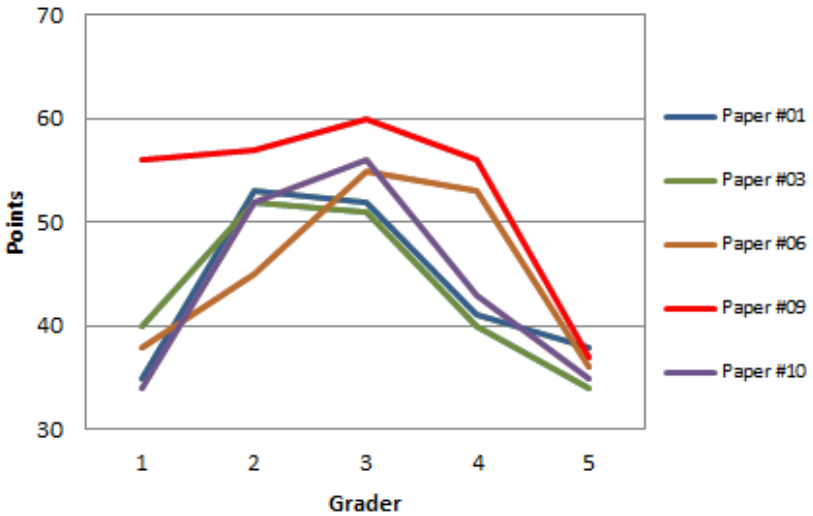

Fig. 3 Points of the firs five works

Table 2 Statistical analysis of the results of the competition in MSc category

No.	Paper	Average	Interval of variation (max-min)	Interval of variation / Average	Average of absolute deviation	Variance	Standard deviation
1	Paper #01	43.8	18	41.10%	6.96	54.16	7.36
2	Paper #02	40.6	22	54.19%	7.12	68.24	8.26
3	Paper #03	43.4	18	41.47%	6.48	48.64	6.97
4	Paper #04	42.6	15	35.21%	4.72	29.84	5.46
5	Paper #05	37.8	9	23.81%	3.44	13.36	3.66
6	Paper #06	45.4	19	41.85%	6.88	58.64	7.66
7	Paper #07	32.2	8	24.84%	2.32	8.96	2.99
8	Paper #08	40.0	14	35.00%	5.20	34.00	5.83
9	Paper #09	53.2	23	43.23%	6.48	67.76	8.23
10	Paper #10	44.0	22	50.00%	8.00	78.00	8.83
11	Paper #11	38.2	20	52.36%	5.52	51.36	7.17
12	Paper #12	34.4	33	95.93%	7.52	110.64	10.52
13	Paper #13	40.0	7	17.50%	2.40	7.60	2.76
14	Paper #14	32.4	12	37.04%	3.68	18.24	4.27

Table 3 The results of the evaluation on students' term papers

Work	Students-graders																			Teacher	
	Grader #1	Grader #2	Grader #3	Grader #4	Grader #5	Grader #6	Grader #7	Grader #8	Grader #9	Grader #10	Grader #11	Grader #12	Sum	Ranking Position	Average	max-min	Variance interval / Average	Average of absolute deviation	Variance	Standard deviation	Grades given by teacher	Ranking Position
#S01	66	57	75	70	70	68	74	71	75	61	74	73	840	**3**	**69.50**	18	25.90%	4.33	29.92	5.58	**60**	**3**
#S02	73	71	73	74	74	45	57	75	74	60	65	74	815	**4**	**67.92**	30	44.17%	7.44	81.24	9.01	**53**	**5**
#S03	62	66	74	68	68	62	70	60	70	68	59	74	801	**6**	**66.75**	15	22.47%	4.13	23.52	4.85	**45**	**6**
#S04	67	70	75	73	75	65	73	70	73	68	68	74	851	**2**	**70.92**	10	14.10%	2.92	10.41	3.23	**61**	**2**
#S05	75	69	75	75	75	68	74	75	75	70	75	64	870	**1**	**72.50**	11	15.17%	3.17	13.08	3.62	**68**	**1**
#S06	67	70	70	69	69	52	75	63	72	63	70	73	813	**5**	**67.75**	23	33.95%	4.33	34.19	5.85	**59**	**4**

This analysis used a relatively small sample. In such contests more numerous commissions rarely appear, and a further analysis of the hypotheses was made on the example of assessing students' term papers. The testing was conducted at the Business school Čačak. The six works (seminar papers), that the students exhibited and defended as part of their regular duties, were a subject to assessment. The evaluation committee consisted of twelve students plus a teacher who evaluated works independently. The results are shown in Table 3

It should be noted that the students-assessors got the auditor's forms with defined grades structure and the span of points for each criterion. This has significantly objectivized ranking and reduced dissipation in the assessment. Such approach was necessary because of assessors' inexperience.

For each work the median was calculated on the basis of evaluations of all assessors. These values are taken as the y variable to calculate the Spearman's rank correlation coefficient on the way described by Weisstein (2012). Individual assessments of each evaluator for the given work are taken for the x variables.

After entering all pairs, a calculation was performed to provide the values of $r_{s(calculated)}$. This gave $r_{s(calculated)} = 0.371$. Since $r_{s(table)} = 0.306$, it is easy to conclude that the $r_{s(table)} < r_{s(calculated)}$. On that basis, according to a Mann-Whitney test (anon, 1999), it is to be concluded that x and y are in correlation. The results were expected because the median is calculated based on the evaluations of all graders.

Grades are matched in only $r_s^{2}{}_{(calculated)} = 13.73\%$, which further indicates that the assessors, the same works under the same conditions, assessed with different grades in the most cases.

This evaluation showed that there were significant differences in scores of individual works. Differences between maximum and minimum of marks awarded to one of works ranged from 10 to 30 points, or one to three grades. However and here it turned out that a group of assessors separated clearly the best and worst work. The best placed work was best placed in nine rankings, and on the remained three lists took the second, fourth and sixth position. The least ranked work was rated on six rankings as the least work, and twice was ranked as the third, fourth or fifth.

The teacher evaluated each work according to the same form that was used by students-graders. Individual grades of papers, which the teacher gave, differed from grades mean values that students-graders assigned to the works, but the overall order of works coincided almost completely. Four works were ranked in the same order in both rankings, but the fourth and the fifth works on the lists had alternate positions. The reason may lie in the fact that the difference between these works on the ranking list formed by the students-graders was only 0.17 points, or 0.17%, so it could be said that these works share the fourth and fifth place. The teacher has noticed a more significant difference between these works.

3 IMPACT OF THE ASSESSMENT SYSTEM ON THE MARK

Previous research has shown that the structuring of the assessment may reduce the impact of the experience of assessors and that comparable results in terms of quality of works can be obtained, but significant differences still occur in the individual assessments of the works. However, both assessments were conducted on the way that all points, assigned by the evaluators, were added for each work, and the final order of ranking was determined by these totals. It is interesting to see if it would give different results with a different grading system.

The new hypotheses can be set. In this case:

Working hypothesis #2

H_0 The grading system may affect the evaluation results.

Alternative hypothesis #2

H_1 The grading system does not affect the evaluation results.

It can be assumed that the criteria of different graders are different, and that they will give different numbers of points to the same works. As noted, for each of the papers, scores were collected, and then, the order of works was determined. That way, the impact of graders who give lower scores, on the final grade, is less than the impact of graders who give bigger numbers of points. In extreme cases, an assessor, who makes a big difference in points between the evaluated works, can dramatically affect the final order.

It is interesting to see what would happen if, to the student competition was applied another system in which every grader had made his ranking, and the points then were tallied based on the position in the tables. Such an analysis was performed and shown in the Table 4. For the purposes of this analysis, the rankings are considered for each grader separately. First place was awarded one point and for each subsequent place one point more. For students who have achieved the same number of points here is awarded the same number of points that corresponded to the better placement.

It is interesting to note that the placement in this system differed from the placement in the original system. Only in seven of the fourteen cases, the placements would remain the same. In one case the difference was three places on the list, three times for one and three times for two places. The students who had the same amount of points, and who shared the same position in the rankings, were assigned with the highest of the rankings among those students (e.g. 8-9 → 8).

Table 4 The results of the evaluation on students' term papers (method 2)

	Work	Grader #1	Grader #2	Grader #3	Grader #4	Grader #5	Sum	Position (way 2)	Position (original)
1	Work #09	1	1	1	1	3	7	1	1
2	Work #01	7	2	5	7	1	22	2	4
3	Work #06	6	7	3	2	5	23	3	3
4	Work #10	8	3	2	6	7	26	4	2
5	Work #04	3	9	7	3	8	30	5	6
6	Work #03	5	3	6	8	8	30	5	5
7	Work #13	2	8	9	9	3	31	7	8-9
8	Work #11	8	3	11	12	1	35	8	10
9	Work #08	10	11	8	4	8	41	9	8-9
10	Work #02	10	13	3	5	11	42	10	7
11	Work #05	3	9	9	10	12	43	11	11
12	Work #12	14	6	12	12	5	49	12	12
13	Work #14	13	12	14	11	12	62	13	13
14	Work #07	12	13	13	14	12	64	14	14

In case of use of some other method for the final rankings such as use of weighting of the rankings similar to F1 (anon, 2012), it is most likely that a ranking list will be different. Thus, for example, if for each position on the list of each grader, the number of points as shown in Table 5 is applied, a new ranking will be established as shown in Table 6 in the column Position (pondered).

Table 5 Weights for the rank positions

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Points	25	20	16	13	11	9	8	7	6	5	4	3	2	1

In case of such a scoring system, in which the first five positions were significantly more valuated, there are only five matches with the original rankings, but, also, only seven matches with the previous ranking list obtained the same way, but without weighting. Five works has achieved the same position on all three rankings!

Specific examples confirm the hypothesis, and, based on the used sample, there is no reason to reject the null hypothesis #2. For the alternative hypothesis #2, in this sample, there are observed only five confirmations, so it could be said that it is very unlikely, and that the alternative hypothesis #2 was not confirmed in this case.

What has the prior analysis shown? Roughly, one might say that the score depends largely on the impression that the student leaves on the examiner. On the other hand, it is shown that truly valuable works, which considerably deviate from the middling, are noticed and well-marked. Among the mediocre works, the scores vary considerably, probably due to different affinities, concentration and the current mood of assessors. In all this, at no time is included in the analysis neither mental nor physical condition of the student, what undoubtedly affects the results achieved on the exam.

Table 6 Comparison of possible outcomes for the three different evaluation methods

	Work	Position (pondered)	Position (method 2)	Position (original)
1	Work #09	1	1	1
2	Work #01	2	2	4
3	Work #06	3	3	3
4	Work #10	4	4	2
7	Work #13	5	7	8-9
8	Work #11	5	8	10
5	Work #04	7	5	6
6	Work #03	8	5	5
10	Work #02	9	10	7
9	Work #08	10	9	8-9
11	Work #05	10	11	11
12	Work #12	12	12	12
13	Work #14	13	13	13
14	Work #07	14	14	14

When both the previous null hypotheses were confirmed, the question about the method to get a realistic measure of knowledge and skills of the students can be justifiably brought up. A possible solution to eliminate the impact of graders on the results of the exam could be the application of computer adaptive testing (CAT).

4 APPLICATION OF THE COMPUTER ADAPTIVE TESTING (CAT) IN THE EVALUATION

Machine testing of knowledge and skills of students was proposed by many authors, and a large number of authors propose CAT as a way to eliminate subjectivity in the assessment and reduction of the time of students testing, such as Curtis (2009), Maravić-Čisar&All (2010), Clariana & Wallace (2002), and others. However, it should be noted that machine testing can be problematic in the case of blind persons testing. Some practical guidelines for carrying out of testing in these cases are presented in the TEST ACCESS: Guidelines for Computer Administered Testing (Allan, Bulla, & Goodman, 2003)

Theoretical bases of CAT applied in this research are presented in the doctoral dissertation of Svetlana Andjelic (2010). In the dissertation, the presented CAT model contains a testing that is divided into different levels of difficulty, similar to Zenisky, Hambleton and Luecht (2010). The idea is that in the preparatory phase of the database creating, a large number of predefined questions, in all areas, were given to a large number of students. On the basis of their responses it is possible to define sets of questions for each area. Questions are grouped into groups by areas and by the severity of responses. When the questions base is formed, students can start testing in accordance with CAT model shown in Figure 4. Student, in accordance with the expected level of his knowledge, chose the level where testing begins. In this way, student can affect the length of the testing, but the selected initial level in any case does not affect the final result of testing.

4.1 Research method

The experiment, consisting of two separate phases: the classical testing and CAT testing, was conducted on a sample of 100 students.

4.1.1 Phase 1

The students took the exam using classical, paper based approach. The questions were of various difficulties and were rated accordingly, each carrying an adequate number of points. To enable an easy grade forming process, the sum of the points from all the questions was 100. All of the students were given 14 identical questions, with 8 questions carrying 5 points, and 6

questions carrying 10 points. Students did not lose points for wrong answers. The students were given one hour to complete the test. After that, the professor evaluated the tests and gave grades to the students. This part of the process took around six hours. In addition to the evaluation of the tests, the professor had to mark each point on the paper, check the sum of points, write it into the chart, check if all the grades were filled in properly etc.

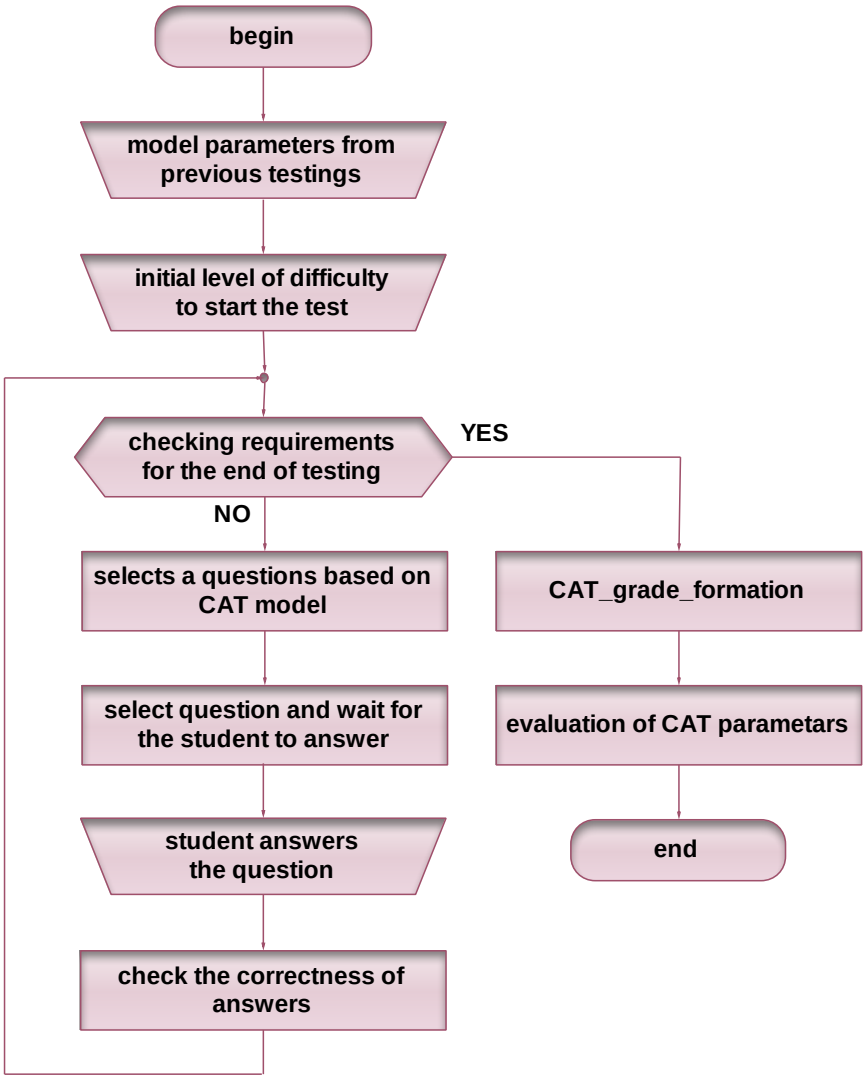

Fig. 4 Algorithm of CAT testing

4.1.2 Phase 2

The same students were tested using a computer. The testing was based on the above described CAT model and conducted in the computer lab. It is worth mentioning that the students got familiar with the software using it during lab classes. This was necessary in order to avoid stress during the testing that could have been caused due to the use of unfamiliar software. As it always happens, the level of the familiarity with the software was not the same for all students.

4.2 The research results

In the following example, questions were divided into three levels of difficulty, and each question is followed by five answers. At the very beginning, probability for the final grade is the same for all possible outcomes because it can be presumed that at the beginning of the

testing all outcomes are equally probable. If the grading system with six different grades (5, 6, 7, 8, 9 and 10; grade 5 is failing grade) is used equal values of 0.1667 (or 1/6) are taken as starting *a priori* probabilities.

For the initial a priori probability of questions the calculated values based on the classic paper test can be taken (Linacre, 2000). The values used as a priori probabilities were arrived at on the basis of the test results from the first phase (the paper based test). After the completion of the CAT testing, a new evaluation of the model was conducted, or in other words, new a priori probabilities were calculated. New probabilities (a posteriori probabilities) are calculated based on Bayes' theorem of conditional probabilities, similar to the proposed method of Rudner (1998) and Rudner & Liang (2002), on which is then applied Maximum a posteriori (MAP) approach (Pavlek, 2005).

Here will be presented an example of the process of grade forming for one student (testing was conducted at the Faculty of business and industrial management – "Union" University Belgrade).

Through an application of the MAP model to the last calculated *a posteriori* probabilities (the probabilities after the last item), it can be seen that it is maximal for the event A_8[1] . Therefore, it can be concluded that the student has earned grade 8. This can be clearly seen on the Fig. 6, which shows a posteriori probabilities after each item of the test. Probabilities are shown for the moment after giving answer to the question. Student answered the question, and answers were true (T) or false (F) as it is shown on the Fig. 5.

It is interesting that after only six questions, the probability of the final mark 8 was almost three times greater than the second the most probable mark, mark 7. After the third answer it was clear that the student will earn mark 7, or mark 8.

The same student was tested by one classical test with 14 questions. He answered on nine of fourteen questions. Five answers were wrong. If a posteriori probabilities were calculated after each answer, the results would look like as it is shown on the figure 6.

This student got a mark 8 for his answers, the same mark in both cases. Comparison of CAT and classical marks for all students gave results as shown on the Fig. 7.

The figure 7 shows that 52% of students earned the same marks under both methods. In remained 30% of cases students earned better marks when they were tested classically, and in 18% of cases they got better grades with CAT method. In 20% of cases the difference was grater then one mark, and in 8% of cases difference was 3 grades. The later detailed analysis showed that the cause of bigger differences laid in the fact that some of students were not well prepared for CAT testing. Smaller differences mainly were consequences of the fact that the marks were rounded and given as integers. In such cases small differences in the area around the half's gave different marks.

In order to test whether there is a connection between each student's grades after the first and the second phase, the Mann – Whitney test was applied. In this particular case, the one - sided hypotheses testing the existence of correlation between two variables without getting into the type of correlation (positive or negative) were used.

Based on the results of the mentioned statistical test, it has been shown that there is a connection between the final grades that student received after each of the tests, conventional and CAT. The grades match in $r_{s\,(calculated)}^{2}$ = 82.57% of cases. The conclusion is logical because it shows that the difference in grades that each student received is small and amounts to one grade at most (which can be verified by inquiry into the grades of each student).

[1] A_i designates events (the outcomes of testing) where i = 5 to 10 (i.e., grades 5 through 10)

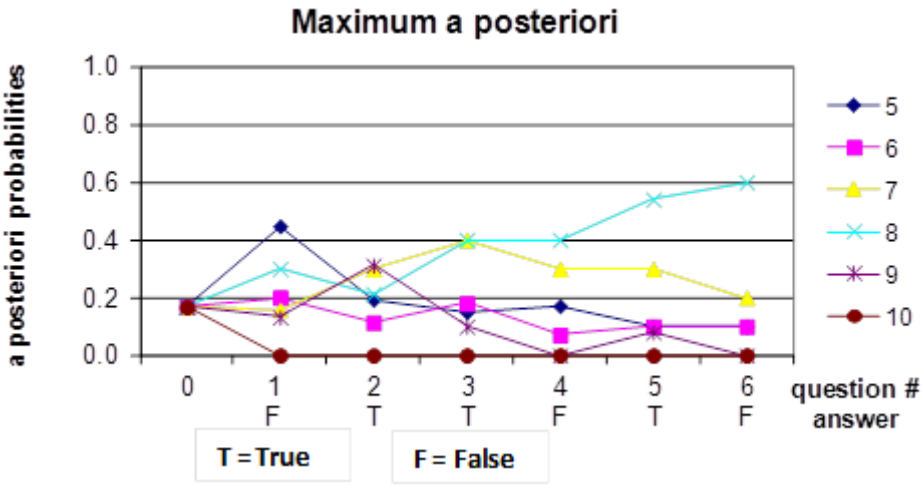

Figure 5 Determination of the mark based on CAT testing

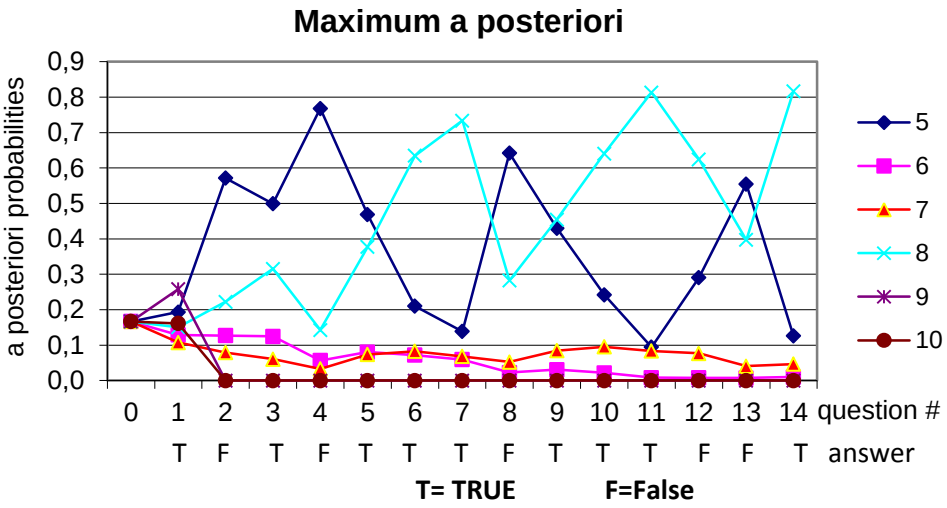

Figure 6 A posteriori probabilities of events in classical testing of student #53

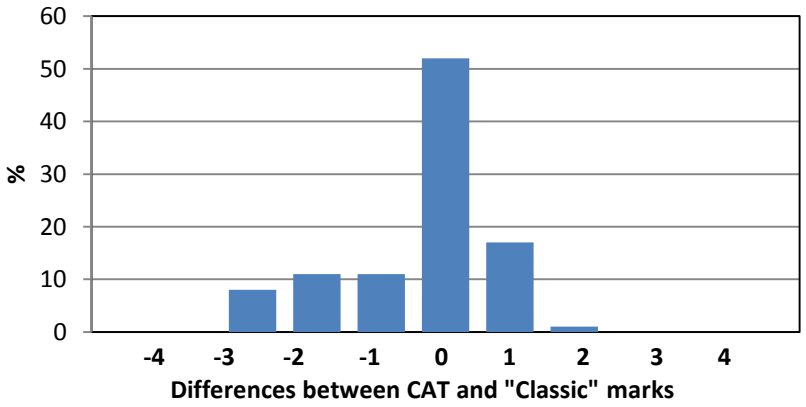

Figure 7 Comparison of CAT and classical marks of students

It is also proved that there is a correlation between the grades that each individual student received after each of the tests, or in other words, the grades are the same, or slightly different (the difference of up to one grade). There is, however, no correlation between the total numbers of specific grades at each test. The total number of specific grades (5, 6, 7...10) received at the classic test is not mutually connected with an adequate value for the CAT test. One of the main reasons for this is in possible mistakes in the choice of answers in the multiple choice questions of the classic paper based test.

After completing the CAT testings, students had to fill in a questionnaire. The survey aimed to summarize the students thoughts and suggestions on the implementation of electronic testing. Special emphasis is put on the CAT testings in education. In line with this the questions in the survey were aimed on the application of computer adaptive testings in the education. Students were reminded that survey results do not affect their grades.

When asked whether they thought that the classic test objectively evaluate their knowledge 23% of surveyed students answered affirmatively.

That traditional testing does not suit them declared 82% of surveyed students, and their answers can be summarized as follows:

- Testing is not adapted to me (35%),
- I am reluctant to answer the questions that require long answers (23%),
- The same answer to a question for different students was scored differently (25%).

Even 70% of surveyed students responded that CAT testing suits to them, with the explanation:

- It is easier to answer questions that offer answers (45%),
- I got neither too easy nor too difficult questions, and it shows that the questions were tailored to me (38% surveyed studenats),
- CAT testing is interesting and not monotonous (32%),
- Teacher can not affect the assessment (20%).

It is interesting that 74% of surveyed students were considered that applied CAT testing objectively reflects their knowledge (Fig. 8).

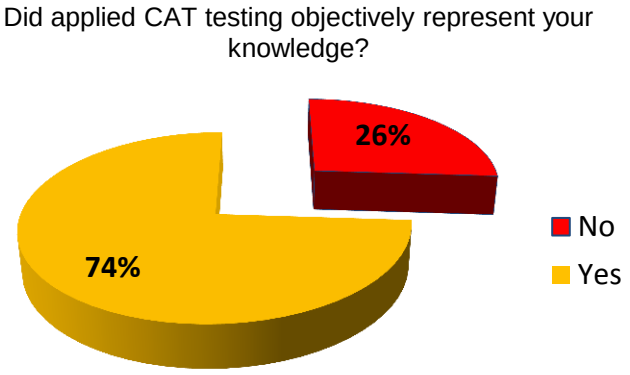

Fig 8 CAT testing and objectiveness of the assessment of knowledge

Students who have declared that they do not like CAT testings are mostly of those who have declared that they use computers rarely, and they belonged to older age (between 30 and 50 year of life). Students stated the following reasons:

- lack of familiarity with computer, even the fear of computers, (10% of students),
- not adapted to interface (5%),
- inability to self-check the obtained results (15%).

5 CONCLUSIONS

In this paper it is shown and proved that each test of knowledge and skills carries a series of

traps and that each score is subject to discussion. It also demonstrates the significant impact of examiners on the final assessment in the case of using of conventional grading system.

So, it is to be concluded that one of the main problems that teachers and students face is the objective evaluation of the level of knowledge in certain field. From the perspective of the teacher, the problem is evaluation of knowledge and description of this evaluation with a grade, as a measure of evaluation of that knowledge. From the perspective of student, it is important that he is given an objective and adequate grade that describes the level of knowledge he has.

As one of solutions that can give an objective grade it can be suggested the use of CAT *Grade forming function* which defines the grade using different approaches, such as Bayes' theorem and MAP procedure. This has been proven to be very efficient in most cases.

Use of a function for determining the conditions for the end of the test allows ending the test when the grade clearly converges toward one of the possible values.

In cases of classic testing and bigger frequency of one question, it has been noticed that students pay more attention to that question and that the percentage of the correct answers increases. This way, the questions that have been labeled as hard remain in a group of hard questions without a good reason. As a consequence the evaluation becomes less objective. CAT testing reduces the amount of these situations. It is possible to provide automatic update after each test and ensure statistically stable data.

The advantages of use of the described CAT model in practice for students and teachers are vast, and much bigger than the disadvantages. The need for mass pretesting, in order to get relevant questions, answers, and distractors (incorrect options in a multiple choice question), is seen as the main disadvantage of the model by the authors. This disadvantage can be eliminated or minimized through a creation of databases of questions, answers, and distractors that would be common for several faculties and/or universities. The work on implementation of this model of estimation of student's knowledge as standard procedure is therefore justified. Having in mind the large number of candidates that attend the lectures from the same subject, CAT can be used as a powerful tool for evaluation of students in high school as well.

Finally, it should be given an answer to the reasonable questions, from the beginning of the paper, about the appropriateness of evaluation: Does evaluation make sense and, if so, whether the mark gives the real picture about the knowledge and skills of interrogated student? On the basis of the above, the answer can be given that assessment makes sense. In the paper conducted research has shown that the evaluation was a very efficient in selection of the best and the worst works, and that it is much harder to distinguish between works of similar quality level. If during the study, in the frame of the teaching subject, several control assessments are performed, it is very likely to get useful information about the level of students' knowledge from that subject, and the final mark for this subject can be pretty realistic. Since the student, during the study, is assessed in many different teaching subjects, it is very likely to be obtained, from all marks, usable information about his capacities. Application of CAT could remove one more problem, lack of uniformity in assessing at the various colleges and universities. This way ranking by the candidate could be done, regardless of where the student earned his degree.

Works Cited

Allan, M. J., Bulla, N., & Goodman, A. S. (2003). *TEST ACCESS: Guidlines for Computer Administered Testing.* Louisvill, KY: American Printing House for the Blind: Louisville, KY.

Andjelic, S. (2010). *A contribution to objective evaluation of students' knowledge useing computer adaptive testing.* Krusevac: Faculty of Industrial Management in Krusevac, "Union" University Belgrade.

anon. (1999). *Interpreting the Mann-Whitney test*. Retrieved 6 8, 2012, from GraphPad Software: http://www.graphpad.com/articles/interpret/analyzing_two_groups/mann_whitney.htm

anon. (2012). *Points.* Retrieved 6 29, 2012, from Formula 1: http://www.formula1.com/inside_f1/rules_and_regulations/sporting_regulations/8681/

Bergen. (2005). *TOWARDS THE EUROPEAN HIGHER EDUCATION AREA.* Retrieved 6 8, 2012, from bologna-bergen2005: http://www.bologna-bergen2005.no/Docs/00-Main_doc/010519PRAGUE_COMMUNIQUE.PDF

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *Acta Polytechnica Hungarica, 33*(5), 593-602.

Curtis, P. (2009, 07 12). *theguardian.* Retrieved 06 07, 2012, from Computerised testing likely to replace traditional exams, says head of board: http://www.guardian.co.uk/education/2009/jul/12/written-exams-computerised-testing

EU. (1999). *The Bologna Declaration.* Retrieved 6 8, 2012, from europedu.org (Sorbonne-Bologna process, French Ministry of Education): http://ec.europa.eu/education/policies/educ/bologna/bologna.pdf

EU. (2001). *TOWARDS THE EUROPEAN HIGHER EDUCATION AREA.* Retrieved 6 8, 2012, from Onderwijs.vlaanderen.be: http://www.ond.vlaanderen.be/hogeronderwijs/bologna/documents/MDC/PRAGUE_COMMUNIQUE.pdf

EU. (2003). *http://www.bmbf.de/pub/communique_bologna-berlin_2003.pdf.* Retrieved 6 8, 2012, from Bologna-Berlin 2003: http://www.bmbf.de/pub/communique_bologna-berlin_2003.pdf

EU. (2007). *London Communiqué.* Retrieved 6 8, 2012, from Onderwijs.vlaanderen.be: http://www.ond.vlaanderen.be/hogeronderwijs/bologna/documents/MDC/London_Communique18May2007.pdf

EU. (2009). *Communiqué of the Conference of European Ministers Responsible for Higher Education, Leuven and Louvain-la-Neuve, 28-29 April 2009.* Retrieved 06 08, 2012, from Onderwijs.vlaanderen.be: http://www.ond.vlaanderen.be/hogeronderwijs/bologna/conference/documents/Leuven_Louvain-la-Neuve_Communiqu%C3%A9_April_2009.pdf

Linacre, J. M. (2000). Computer-Adaptive Testing: A Methodology Whose Time Has Come. *MESA Memorandum*(69).

Maravić-Čisar, S., Radosav, D., Markoski, B., Pinter, R., & Čisar, P. (2010). Computer Adaptive Testing of Student Knowledge. *Acta Polytechnica Hungarica, 7*(4), 139-152.

Pavlek, S. B. (2005). *Bayes' learning.* Retrieved 6 7, 2012, from The Institute of Electronics, Microelectronics, Computer and Intelligent Systems: http://www.zemris.fer.hr/education/ml/nastava/ag20022003/bayesovo_ucenje.ppt

Rudner, L. M. (1998). *An On-line, Interactive, Computer Adaptive Testing Mini-Tutorial.* O'Boyle Hall, Washington, DC: ERIC Clearinghouse on Assessment and Evaluation.

Rudner, L. M., & Liang, T. (2002, 06). Automated Essay Scoring Using Bayes' Theorem. *The Journal of Technology, Learning, and Assessment, 1*(2), 3-21.

Weisstein, E. W. (2012). *Spearman Rank Correlation Coefficient.* Retrieved 06 29, 2012, from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/SpearmanRankCorrelationCoefficient.html

Zenisky, A. H. (2010). Elements of Adaptive Testing. Statistics for Social and Behavioral Sciences. *Multistage Testing: Issues, Designs, and Research.*, 355-372.

Received for publication: 15.12.2012
Revision received: 14.02.2013
Accepted for publication: 29.03.2013

How to cite this article?

Style – **APA** *Sixth Edition:*

Čekerevac, Z., Anđelić, S., & Čekerevac, P. (2013, 07 15). Knowledge assessment and application of computer adaptive testing. (Z. Čekerevac, Ур.) *MEST Journal, 1*(2), 16-30. doi:10.12709/mest.01.01.02.02

Style – **Chicago** *Fifteenth Edition:*

Čekerevac, Zoran, Svetlana Anđelić, / Petar Čekerevac. „Knowledge assessment and application of computer adaptive testing.“ Уредник Zoran Čekerevac. *MEST Journal* (MESTE) 1, бр. 2 (07 2013): 16-30.

Style – **GOST** *Name Sort:*

Čekerevac Zoran, Anđelić Svetlana and Čekerevac Petar Knowledge assessment and application of computer adaptive testing [Journal] = Knowledge assessment and application of CAT // MEST Journal / ed. Čekerevac Zoran. - Belgrade : MESTE, 07 15, 2013. - 2 : Vol. 1. - pp. 16-30. - ISSN 2334-7058 (Online); ISSN 2334-7171.

Style – **Harvard** *Anglia:*

Čekerevac, Z., Anđelić, S. & Čekerevac, P., 2013. Knowledge assessment and application of computer adaptive testing. *MEST Journal,* 15 07, 1(2), pp. 16-30.

Style – **ISO 690** *Numerical Reference:*

Knowledge assessment and application of computer adaptive testing. **Čekerevac, Zoran, Anđelić, Svetlana and Čekerevac, Petar.** [ed.] Zoran Čekerevac. 2, Belgrade : MESTE, 07 15, 2013, MEST Journal, Vol. 1, pp. 16-30. ISSN 2334-7058 (Online); ISSN 2334-7171.