Comparison of Examination Methods Based on Multiple-choice Questions Using Personal Computers and Paper-based Testing

[1]Sanja Maravić Čisar, [2]Robert Pinter, [3]Dragica Radosav, [4]Petar Čisar

[1]Subotica Tech-College of Applied Sciences, Subotica, sanjam@vts.su.ac.rs,

[2]Subotica Tech-College of Applied Sciences, Subotica, probi@vts.su.sc.rs

[3]Technical Faculty "Mihajlo Pupin" Zrenjanin, University of Novi Sad, radosav@tfzr.uns.ac.rs

[4]Telekom Srbija, Subotica, petar.cisar@gmail.com

Contribution to the State of the Art

DOI: **10.7251/JIT1201022C** UDC: **004.032.6**

Abstract: Computer-based testing, by facilitating the interaction between teaching and learning, can improve the quality of learning through improved formative feedback which is a key aspect of formative assessment. This study makes a contribution to the research on computer-based testing by examining the mode differences between the paper-and-pencil test and computer-based test. The previously conducted researches in this area dealt with the students of primary and secondary schools. In those researches the points of observation were the students' successes in mathematics, English and social sciences; no research was done in field of programming languages such as C++ with post-secondary students.

The main aim of this study was to find out whether there are differences in the achieved results in two ways of testing: computer-based testing and paper-and-pencil test. Also, the intention was to detect those characteristics of computer based test, which may have a negative effect on students' achievements. The participants were a representative sample of the population of all engineering students studying computer science at Subotica Tech. The findings of this study led the authors to reach the conclusion that there are no significant differences in scored results for the paper-and-pencil testing and the computer-based testing.

Keywords: computer-based test; paper-and-pencil test; assessment; testing; post-secondary education

Introduction

Traditional methods of assessment have limited capabilities in measuring the learning and progress of each student, especially in guiding the study process. These methods are particularly inappropriate today, when knowledge and the working environment change rapidly and complement each other, and the ability for independent lifelong learning is becoming more than necessary.

Modern technology offers many possibilities for improving the process of education and knowledge assessment. The history of using computers to perform the review process of knowledge begins with the 1970s [6]. However, the high price of computers at that time and their technical capabilities limited their application for testing. The progress of technology enabled the development and application of computers for testing in many areas, including the education process.

In the system of education, testing and evaluation of knowledge is of particular importance. Checking and evaluating knowledge enables teachers to determine the level to which students adopted the curricula and gained some knowledge and to get feedback about their work and applied teaching methods in order to improve it.

The marks are described as quantitative, numerical, qualitative, i.e. descriptive and by ranking or analytical. The criteria for evaluating the success of students are type, scope and level of approved knowledge, and skills in relation to what is prescribed by the curriculum of post-secondary institutions. In order to test whether evaluation has the proper effect, it is of great importance for the teacher's assessment of student knowledge to be accurate, objective and reliable.

The true strength of assessment is reflected in the feedback information to students. Improving the quality of the learning process involves not only the final determination of student knowledge at the end of the course, but more importantly the measurement of achieved knowledge during the course. Thereby students are more strongly motivated by their success in learning, they are taking more self-responsibility in the process of learning, they discover their "strong and weak points", and thus become active participants in learning.

The wide-spread popularity of computers resulted in directing attention to the possible use of computers in the process of knowledge evaluation. Advantages and benefits of this method of assessment and knowledge evaluation are various: the time needed to review the work of students is significantly reduced, there is the possibility of statistical analysis of questions, cost reduction in comparison with the validation of knowledge which includes printing tasks, the application of multimedia in setting questions, the possibility of measuring the time needed for response, and increasing the level of security.

However, all these advantages of computer-based testing become irrelevant, if it turns out that the test of knowledge with computers has side effects for individuals, i.e. it is not appropriate for all students.

Since there is an increasing number of schools in Serbia that have PC laboratory rooms, there is a growing interest in computer based assessments. However, there is also the ever-present question of the value and comparability of the results that are attained on computer tests and in the conventional way. The primary concern is whether the form of test delivery affects the results achieved by students on the test. For example, it is possible that the level of skills in computer use affects the final result of the test when compared with the result of the same test but in paper format.

The research that was done has an empirical - theoretical character. The problem into which the research was conducted was to investigate, whether the delivery of knowledge (computer or paper-and-pencil test) in the process of evaluation has a statistically significant impact on the results achieved and in increasing the quality of the teaching process. Following the research, students completed the questionnaire about their attitudes towards this kind of knowledge testing, in what way and whether the manner of presenting questions (one question or several questions simultaneously shown) had any impact on the achieved results.

Literature review

The use of computers in the process of testing began in the early 1970s. Initially, the technical capabilities of computers and their prices restricted the use of computerized tests. With the advantages that the new technology provides, this type of testing is beginning to develop, and consequently there are a number of researches that examine the role and application of computers in the process of knowledge evaluation.

According to the Guidelines for Computer-Based Tests and Interpretations from American Psychological Association (APA) [2], score comparability or equivalence between computer-based tests and paper-based tests is defined as follows: "Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode."

Lee and Hopkins [11] in their study found that the mean paper-and-pencil test score was significantly higher than the mean computerized test score. They also concluded that only software that allows the conveniences of paper-and pencil test, e.g., the ability to change answers and the ability to review past items, should be used in future applications.

The study of Shermis and Lombard [16] examined the degree to which computer and test anxiety had a predictive role in performance across three computer-administered placement tests (math, reading, written English). Results showed that age and test anxiety were both significant predictors for math performance, with lower values on the two variables associated with better performance. When reading was the outcome variable, age and computer anxiety were statistically significant performance predictors, with older readers faring better and less anxious individuals achieving higher scores. No predictors were statistically significant for the written English essay.

Nichols and Kirkpatrick [15] explored the impact of the mode of presenting the test for the Florida state assessment in high school reading and mathematics. They found that for both reading and mathematics, the mean raw score, mean scale scores, and passing rates were slightly higher for paper-and-pencil test (PPT) than for computer-based test (CBT), although the mode effect was not significant.

Way et al. [19], investigated the comparability of paper and online versions of the Texas statewide tests in mathematics, reading/English language arts, science and social studies at grades 8 and 11. The results of this study showed that the tests were more difficult for the online group than for the paper group.

Keng et al. [9] found that English language arts items that were longer in passage length and math items that required graphing and geometric manipulations or involved scrolling in the online administration tended to favor the paper group.

Over the years, the quality of tests that are done on the computer has changed, also the student experience in using computers. The study of Kingston [10] summarizes the results of eighty-one researches that have been done between 1997 and 2007. All these studies investigated the comparability of classical test and test done on computer. In his study, Kingston applied meta-analysis in order to demonstrate if the grade (elementary, middle or high schools) or subject in which knowledge is checked (English, mathematics, social sciences) have an impact on the comparability of computerized and traditional tests. Research has shown that the grade does not affect the comparability of tests, while in the case of the subject it was shown that the classical tests have a small advantage for math test, while a computerized test of knowledge has an advantage in testing English and social sciences.

The paper of Wang [18] described the research that was done in 2003 in the United States. The subject of study was Stanford Diagnostic Reading Test Fourth Edition (SDRT 4) and the Stanford Diagnostic Mathematics Test Fourth Edition (SDMT 4), each of which has six levels and which are adapted for taking on the computer. The participants were students from U.S. school from second to twelfth grade. In this study, 1863 students have done the test SDRT 4 and 1774 students the test SDMT 4. The results gave solid, unambiguous evidence of reliability and comparability of test results SDRT 4 and SDMT 4 for all grades and levels of the test, regardless of the manner of conducting the test. Differences in the achieved results based on the method of conducting the test do not exceed the expected random errors for most SDRT 4/SDMT 4 subtests.

The project PASS-IT (Project on Assessment in Scotland – using Information Technology) lasted for 27 months (starting from August 2002 until December 2004), its aim was to look into the possibility of formative and summative online knowledge assessment in secondary schools in Scotland [3]. One of the conclusions of the research is that technology must support the educational requirements of specific subjects and levels. For example, in order to reliably and validly determine the success of students in mathematics, the system must provide the possibility of partial points. Furthermore, for certain subjects such as music, integration of multimedia elements is very important to support the issues in this area.

Today's technology has the ability to do more than just accelerate the process of testing. A growing number of experts involved in education agree that technology can improve teaching and learning. One of the projects that involves new forms of technology in solving problems in real life is the Problem Solving in Technology Rich Environments (TRE project) [17]. The project was started in 2003 in the United States and had a number of participants of 2000 students. TRE tested necessary scientific skills, such as the ability to find information about preset subject, to estimate which information is relevant for experiment, to make the plan and perform an experiment, and to organize and interpret results.

Thus, for example, eighth grade students in the experiment (which was entirely done on computer) had the task of using a balloon charged with helium to solve the problem of the growing complexity. They had to find the relation between power holding balloon at a height, mass and volume. Students were asked to determine the relationship between the mass which is placed in the basket of balloon and height it can reach. To solve this problem, students have gathered the necessary information performing the experiment several times with different masses, and when they had enough data to make conclusions, it was supposed to give the conclusions in the form of answers to multiple responses questions. The TRE project demonstrated several unique capabilities of knowledge assessment provided by technology [17]. First, the technology allows the presentation of much more complex problems to be solved in several steps. Different forms of multimedia, such as an animated helium balloon and an instrument panel that allows setting the parameters of the balloons, can represent the problem much better than if it were only explained in written form or orally.

Another example of technology in setting up and solving problems is the Floaters test which is offered to students in the UK as part of the World Class test [17]. This program allows checking students' knowledge in conditions without paper and pencil. For example, students use interactive simulation to measure the weight of various foods such as carrots, apples and bananas, and their task is to determine whether these pieces of fruit can float on the surface of the water. Students are then asked to set up a hypothesis based on the templates that were found.

Research

The main purpose of this study is based on theoretical research and the use of computer capabilities in the evaluation of knowledge in order to indicate statistically significant possibility of raising the overall level and quality of the teaching process. Some results about using computers for student assessments could be found in Maravic et al. [12] and Maravic et al. [13]. Besides this main purpose, the aim was to detect those characteristics of CBT, which may have a negative effect on students' achievements. The objective was to determine the influence of the way in which computer randomly generates questions (area and weight), i.e. an impact if first the most difficult question appears from a set of selected test questions and inability of browsing back and forth. Also, the intention was to find out if there is influence on students' results if immediately after given the answer the message "answer is correct" or "answer is incorrect" appears.

The main hypothesis of this research is that the results, given by the computer-based tests, are valid and reliable alternative to the classical way of knowledge testing on paper. Therefore, the goal is to find the answer to the question of whether there are differences in the achievements of students which outcome from different modalities of delivery of the test. The following null hypothesis was stated:

"There is no significant difference between the students' score in computer-based test, compared to those obtained with paper-and-pencil test."

In addition to this primary aim, one more objective was formulated: how and whether the way of question presentation (one question at a time on the screen, or more questions and need for scrolling) affects achieved better result. The following auxiliary hypothesis was stated:

"There are differences between the students' score which depend on how the computer-based assessment was built, how the question was presented and which are the answering techniques."

Experimental Results

The experiment was carried out with college students. The objectives were to evaluate students' results and opinion when they take tests on PCs, to see whether there are significant differences in the results obtained with paper-and-pencil tests and with computers, and also to search for differences between diversely assembled computer based assessments. In order to know what the students' opinions are and whether or not they were satisfied, a survey was carried out with specific questions and personal comments.

Participants

The participants were future engineers, i.e. students of computer science (engineering students) at Subotica Tech (Serbia), all about the same age (about 20 years old) and in a similar situation (first year of computer science study). Data was collected in the spring of 2010. The research included 90 students (selected from the Department of Informatics) who took the Object-oriented programming course as a compulsory subject. The students of computer science are predominantly male (which is generally true for Subotica Tech). This is reflected in the gender-percentage: 90% male test subjects and 10% female. The total number of college students at Subotica Tech is 591, of which only 57 are female, or 9.64%, so the sample can be considered representative. The students were divided into two groups, an experimental group with 45 students (computer-based test) and a control group with same number of students (paper-and-pencil test). Students were pre-tested to ensure that the groups are of equal knowledge. All students had previously been given instructions for the examination and related learning material.

Instrument

In order to investigate students' knowledge, a *multiple choice questionnaire* (*MCQ*) with twenty questions was developed. The paper and pencil and the computer based versions of the MCQ test included the same set of twenty questions. The time provided for solving the test was thirty minutes for both groups. Hand scoring was done for the paper-and-pencil version of the test, and automatic scoring by computer for the computer-based test. To make participants familiar with the CBT, they had an opportunity to exercise before the test.

Examination procedure and scoring methodology

All students, participants in the experiment did the same test. The test contained twenty questions. For each question there were several answers offered (usually four) of which only one was correct (i.e. it was an MCQ with one correct answer). Each correct answer carries one point. For incorrect responses there were no negative points given, and questions that remained without answers carried zero points. The negative marking was omitted based on the findings of Bliss [4], namely that negative marking tends to penalize the more able students. The decision to omit questions is influenced by personality characteristics [8]. According to [7], The Royal College of General Practitioners in the UK discontinued negative marking many years ago when they demonstrated that it discriminated female candidates because they tended to be more cautious with regard to guessing.

The maximum number of points that can be obtained on the test was twenty. During the preparation of the paper test, the order of test questions was not associated with their weight (i.e., the questions were not ranged from easier towards the more difficult, or vice versa), but they were randomly selected from a set of questions and compiled to make up the questionnaire. The order of questions for the CBT was left up to the computer to randomly arrange them. Before any of the students used the tests on the computer, all college computers underwent technical checks, to ensure that they had the correct software installed and to check that their display configurations were acceptable. Immediately prior to the test administration, students were asked to access a practice test and practice the question answer submission process.

Analysis

The results that the students achieved on the tests were subjected to statistical analysis. ANOVA (Analysis Of Variance between groups) analysis was applied to test the hypotheses. All statistical analyses reported in this research were conducted with a significant level of .01.

Results

The American Psychological Association (APA), in a document entitled Guidelines for Computer-Based Tests and Interpretations [2], gives specific recommendations for computerized test administrations and score interpretations. The guidelines state the “computerized administration normally should provide test takers with at least the same degree of feedback and editorial control regarding their responses that they would experience in traditional testing formats” [2]. This means that test participants should be able to review their responses to previous items as well as skip ahead to future items, and make any changes they wish along the way. To check the influence of ways of presenting issues two experiments were planned in this research.

In order to check whether or not the way in which questions are presented on the screen may influence achieved results (only one question per screen, or all questions provided for the test), we have carried out two experiments. In the first experiment, participants of the experimental group could only see one question on the computer monitor, there was no possibility of browsing back and forth if an answer to the question was not given, and immediately after submitting the answer the message “true” or “false” appeared on the screen. These factors were obviously available to students of control group who did the PP test.

The distribution of participants’ scores in the PPT and CBT is presented in Table 1 and in Figure 1. The mean score was higher for the paper-and-pencil test (M=9.91, SD=5.22) than for the computer-based testing (M=8.84, SD=4.607) by 1.07 points. The goal of this research was to find out whether there are differences in the achievements of students due to the different modalities of the test delivery. The participants’ results were not statistically different in the CBT and in the P&P test (F=1.056, p>0.01), as presented in Table 2, in the case when students could see only one question on the computer screen.

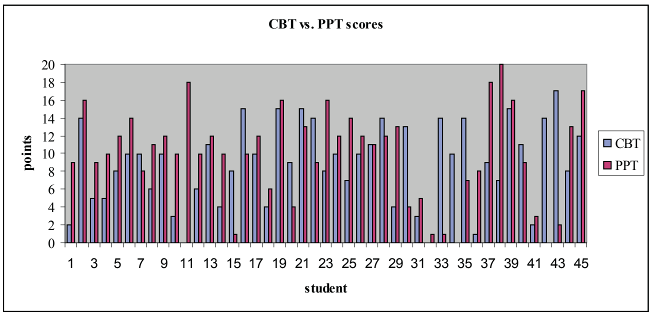

Figure 1. Distributions of students’ scores in PPT and CBT, first experiment

Table 1. The distribution of students’ scores in the PP test and in the CBT, first experiment

	N	Mean	Standard deviation	Standard error	95%Confidence level		Minimum	Maximum
					Lower bound	Upper bound		
PPT	45	9.91	5.222	0.778	9.13	10.69	0	20
CBT	45	8.84	4.607	0.687	8.15	9.53	0	17
Total	**90**	**9.38**	**4.925**	**0.519**	**8.86**	**9.89**	**0**	**20**

Table 2. One-way ANOVA comparison of scores of participants in the PP and CBT, first experiment

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	25.6	1	25.6	1.056	0.306971	6.932
Within Groups	2133.56	88	24.245			
Total	**2159.16**	**89**				

Table 3. The distribution of students' scores in the paper-and-pencil test and in the computer-based test, second experiment

	N	Mean	Standard deviation	Standard error	95%Confidence level		Minimum	Maximum
					Lower bound	Upper bound		
PPT	45	8.93	5.167	0.77	8.15	9.71	0	19
CBT	45	10	5.117	0.76	9.23	10.77	1	20
Total	**90**	**9.47**	**5.141**	**0.542**	**8.93**	**10.0**	**0**	**20**

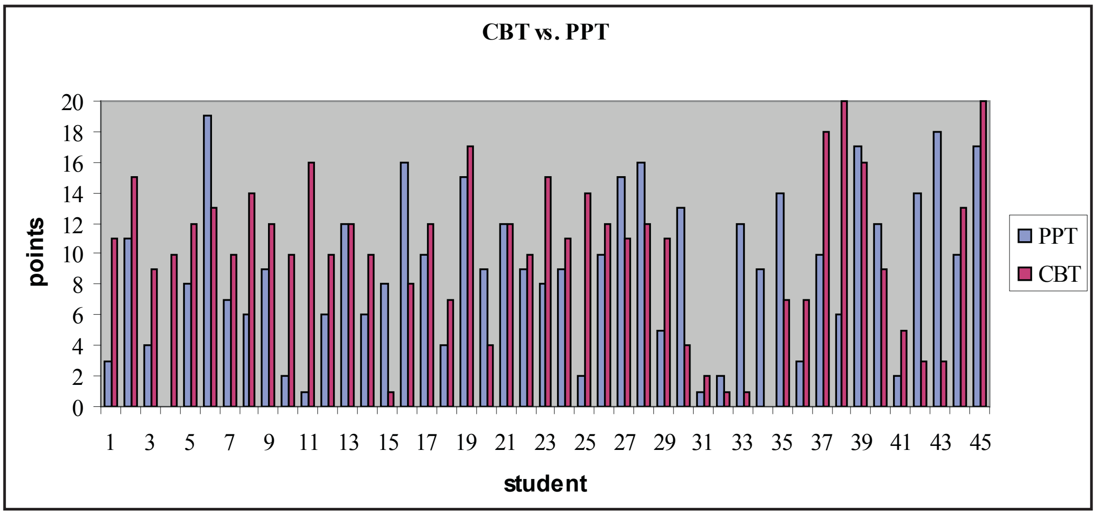

Figure 2. Distributions of students' scores in PPT and CBT, second experiment

To check whether there is a statistically significant difference in the results (which would be the result of the different display modes) the second experiment was conducted two months after the first one. In the second experiment, students who did the test on the computer could see all the questions included in the test at once. After submitting the answer the message "true" or "false" did not appear. The distribution of participants' scores in the PPT and CBT is presented in Table 3 and Figure 2. This time, the mean score and standard deviation for computer-based testing was M=10, SD=5.117, and for the paper-and-pencil test it was M=8.93, SD=5.167. The difference in mean value was the same as in the first experiment, i.e. 1.07 points, but this time students in the experimental group scored better. Data analyses found that there was no statistically significant difference in the results in the CBT and in the PPT (F=0.968, p>0.01), as presented in Table 4.

Based on the results of the first and second experiment we can conclude that there is no statistically significant influence on the students' results due to the way in which questions are presented on the computer screen. The null hypothesis *"There is no significant difference between the students' score in computer-based test, compared to those obtained with paper-and-pencil test"* is confirmed.

After the test, students who did the test on the computer filled out the questionnaire to see what their attitude towards this kind of knowledge testing was, and to find out the answer to the auxiliary hypothesis. The survey was anonymous in order to at-

Table 4. One-way ANOVA comparison of scores of participants in the PPT and CBT, second experiment

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	25.6	1	25.6	0.968197	0.327828	6.931941419
Within Groups	2326.8	88	26.44090909			
Total	**2352.4**	**89**				

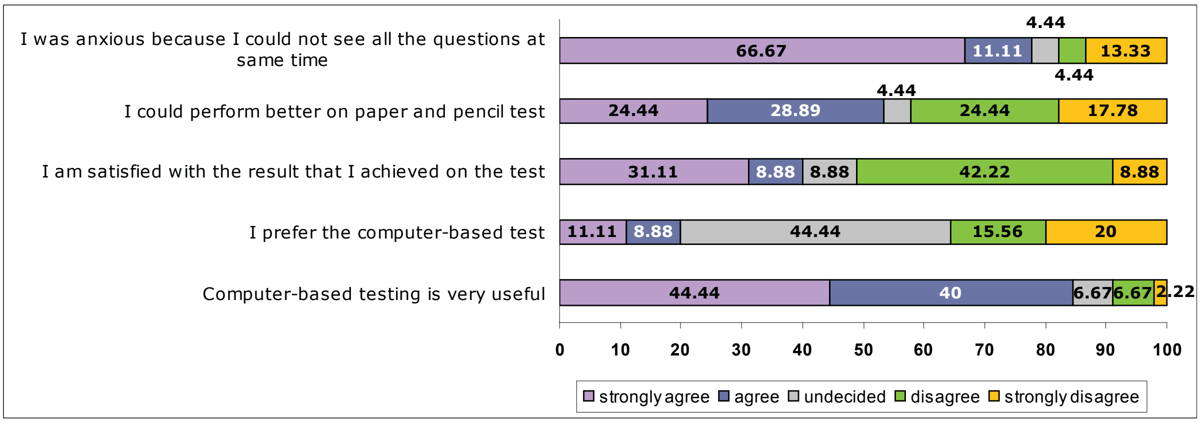

Figure 3. The results of the survey

tain honest answers from the participants. The results of the students' answers are given in Figure 3. The questionnaire was designed to collect information about students' attitudes towards aspects of testing. The survey had five statements: "Computer-based testing is very useful"; "I prefer the computer-based test"; "I am satisfied with the result that I achieved on the test"; "I could perform better on paper-and-pencil test"; "I was anxious because I could not see all the questions at same time". For the evaluation of student responses, the authors used a Likert-type scale with five responses: "strongly agree", "agree", "undecided", "disagree" and "strongly disagree" [5].

Students could also write their personal comments about this kind of testing. The main objection on the part of the students was that they could not see all the questions at once and so could not make a strategy for solving the test. This comment is visible also in their answer to the question "I was anxious because I could not see all the questions at same time", where 66.67% of the students strongly agree and 11.11% agree. Many of the students felt discouraged by the fact that the questions at the beginning of the test seemed too difficult for them, and they would opt for answering them randomly just to get to the next question in line. Later, they had no opportunity to review the test and maybe make an effort to answer the questions that remained unanswered. This attitude may explain the fact that 53.33% of students think that they could perform better on paper-and-pencil test (24.44% strongly agree and 28.89% agree).

As for the results to the statement "I prefer the computer-based test", 11.11% strongly agree, 8.88% agree and even 44.44% were undecided. Students emphasized that they prefer the classical method of solving the test because it gives insight into all the questions for the test. Also, one student "admitted" that he is trying to find a pattern, for example, that the correct answer to every question is under the number 3, and with computer test seeking for patterns was difficult. Despite all the negative comments that were given after the first experiment, students agree that computer-based testing is very useful (44.44% strongly agree and 40% agree). As for the benefits of computer testing, the majority of students pointed out that they liked the fact that after pressing the "submit" button they would find out the result of their achievements. The feedback information after each response about the answer's correctness ("correct" or "incorrect" answer) has a motivational role, but sometimes information that the given answer was the wrong one can negatively affect the further process of solving the test.

After the second experiment, when the students of the experimental group could see all the questions at once, they gave favorable comments. This time the questionnaire had only three statements: "I prefer the computer-based test"; "I am satisfied with the result that I achieved on the test"; "I could perform better on the paper-and-pencil test". The results are given in Figure 4. Students expressed satisfaction because the test now was "a copy of paper test only on the computer". As to the argument of "why" the comments were the following: "I type faster on the keyboard, than I am writing with the pen", "I am more used to use the keyboard than the pencil".

According to the results of the survey, it could be concluded that the following auxiliary hypothesis:

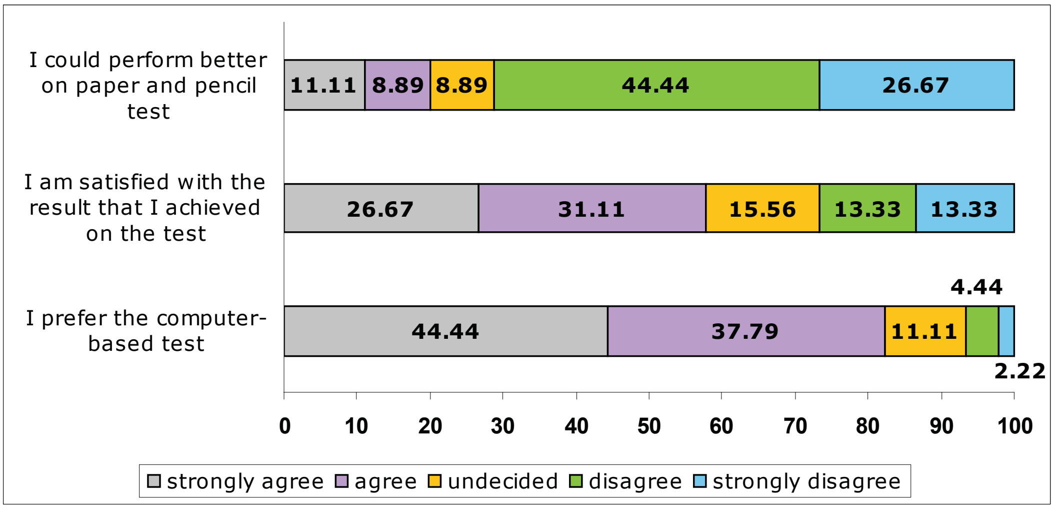

Figure 4. The results of the survey after the second experiment

"There are differences between the students' score which depends on how the computer based assessment was built, how the question was presented and which are the answering techniques." is confirmed.

Based on the presented results, it can be concluded that the advances in computer technology and investments in evaluation and testing software, together with the advantages of immediate feedback and automatic grading, make computer-based testing more and more common.

Conclusions

This study makes a contribution to the research on computer-based testing by examining the mode differences between the paper-and-pencil test and computer-based test. The previously conducted researches in this area dealt with the students of primary and secondary schools. In those researches the points of observation were the students' successes in mathematics, English and social sciences; no research was done in the field of programming languages such as C++ with post-secondary students. Also, the majority of studies were conducted with students in highly developed countries like USA and UK. There are only few studies, for example Akdemir and Oguz [1], which were conducted in a developing country such as Serbia.

The main aim of this study was to find out whether there are differences in the achieved results in two ways of testing: computer-based testing and paper-and-pencil test. Also, the intention was to detect those characteristics of CBT, which may have a negative effect on students' achievements. The objective was to determine the influence of the way in which computer random generates questions (area and weight), i.e. an effect if first the most difficult question appears from a set of selected test questions and the inability of browsing back and forth. The intention was also to find out if it will influence the students' results if immediately after giving the answer, the message "your answer is correct" or "your answer is incorrect" appears on the screen. The participants were a representative sample of the population of all engineering students studying computer science at Subotica Tech. The findings of this study led the authors to reach the conclusion that there are no significant differences in scored results for the PPT and CBT. Also, based on the survey results it can be concluded that the way in which questions are presented on the computer screen does have an effect on student satisfaction with CBT.

It is important to mention that the students were more satisfied with the computer-based test when they could see all questions at once (as in the second experiment). In his study Marks [14] observed that algorithms that randomize the order in which the test questions are presented to each candidate automatically control certain computer-based test assessments. If the test was such that in random sequences first the toughest question appeared, it may increase test anxiety for some candidates and influence their scores. Increased anxiety for whatever reason is likely to have a negative effect on that person's performance on the test.

The answer to this problem could be a computer-adaptive test (CAT), as a form of computer-assisted assessment where the level of difficulty of the questions administered to individual test-takers is dynamically tailored to their proficiency levels. Therefore, a logical continuation of this study is to examine the possibilities and advantages that CAT offers.

References:

[1] Akdemir O and Oguz A (2008) Computer based testing: An alternative for the assessment of Turkish undergraduate students. Computers&Education, 51(3), 1198-1204

[2] American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessment (1986) Guidelines for computer-based tests and interpretations. Washington, DC.

[3] Ashton H and Wood C (2006) Use of Online Assessment to Enhance Teaching and Learning: the PASS-IT Project. European Educational Research Journal, 5(2), 122-130

[4] Bliss L (1980) A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 17(2), 147–152

[5] Clason DL and Dormody T J(1994) Analyzing Data Measured by Individual Likert-Type Items. Journal of Agricultural Education, 35(4), 31-35

[6] Drasgow F (2002) The Work Ahead: A sychometric infrastructure for computerized adaptive tests. In C.N. Mills, M.T. Potenza, J.J. Fremer, & W.C. Ward (Eds.), Computer-based testing: Building the foundation for future assessments, Hillsdale, NJ: Lawrence Erlbaum, 67–88

[7] Goldik Z (2008) Abandoning negative marking, European Journal of Anaesthesiology, 25(5), 349-351

[8] Harden RM et al (1976) Multiple choice questions: to guess or not to guess. Medical Education, 10, 27–32. doi:10.1111/j.1365-2923.1976.tb00527.x

[9] Keng L et al. (2008) Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills, Applied Measurement in Education, 21 (3), 207-26

[10] Kingston N (2009) Comparability of Computer-and Paper-Administered Multiple-Choice Tests for K-12 Populations: A Synthesis, Applied Measurement in Education, 22(1), 22-37

[11] Lee JA and Hopkins L (1985) The effects of training on computerized aptitude test performance and anxiety. Paper presented at the annual meeting of the Eastern Psychological Association. Boston, MA, 1985. Available at http: http://www.eric.ed.gov/PDFS/ED263889.pdf

[12] Maravić Čisar S et al. (2009) True/false Questions Analysis Using Computerized Certainty-based Marking Tests, 7th International Symposium on Intelligent Systems and Informatics SISY 2009, Subotica, Serbia, September 25-26, 2009., Proceedings CD ROM, IEEE Catalog Number: CFP0984C-CDR, ISBN: 978-1-4244-5349-8, Library of Congress: 2009909575

[13] Maravić Čisar S et al. (2010) New Possibilities for Assessment through the Use of Computer Based Testing, 8th International Symposium on Intelligent Systems and Informatics SISY 2010, Subotica, Serbia, September 10-11, 2010., Proceedings CD ROM, IEEE Catalog Number: CFP1084C-CDR, ISBN: 978-1-4244-7395-3, 149-152

[14] Marks AM (2007) Random question sequencing in computer-based testing (CBT) assessments and its effect on individual student performance. Available at http://upetd.up.ac.za/thesis/available/etd-06042008-083644/unrestricted/dissertation.pdf.

[15] Nichols P and Kirkpatrick R (2005) Comparability of the computer administered tests with existing paper-and-pencil tests in reading and mathematics tests. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada

[16] Shermis MD and Lombard D (1997) *Effects of computer-based test administrations on test anxiety and performance*, Computers in Human Behavior, 14(1), 111-123

[17] Tucker B (2009) Beyond the Bubble: Technology and the Future of Student Assessment, Education Sector Reports. Available at http://www.educationsector.org/usr_doc/Beyond_the_Bubble.pdf

[18] Wang S (2004) Online or Paper: Does Delivery Affect Results? Administration Mode Comparability Study for Stanford Diagnostic Reading and Mathematics Tests, Pearson Education, 2004. Available at http://www.pearsonassessments.com/NR/rdonlyres/D381C2EA-18A6-4B52-A5DC-DD0CEC3B0D40/0/OnlineorPaper.pdf

[19] Way WD et al. (2006) Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA, 2006. Available at http://www.pearsonedmeasurement.com/downloads/ conference/ScoreCompTAKS_cp0601.pdf

Submitted: April 15, 2012
Accepted: May 22, 2012