On-line vs. written multiple-choice questions tests: accuracy and usefulness

Tomislav Petković, Zvonko Kostanjčar and Ana Sović
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Hrvatska
Email: sis@zesoi.fer.hr

Abstract— **With the on-line resources becoming common one can ask what are the limitations when applied to on-line examinations. Standardized multiple-choice questions test are commonly used either as a tool to enable student self-examination or as a tool to test a large number of students in a more efficient way. However, the means of delivery and what one intends the tests to measure should be carefully considered. During last two years in teaching the course "Signals and System" both on-line and written, proctored and non-proctored multiple-choice questions test were used in production environment (in vivo). Obtained data suggest that test delivery method strongly influences test accuracy when measuring the knowledge students gained. On-line non-proctored delivery, although being the simplest, yields strongly biased results and is suitable only as a tool to enable student self-assessment of the knowledge gained, while the on-line or written (non-strictly) proctored tests are still the method of choice is ones wants to objectively measure the knowledge students gained.**

I. Introduction

Students have a tendency to learn what teachers inspect rather than what they expect. (Angelo and Cross, 1988 [1]).

Instruction and teaching takes central place in almost all relevant discussions about the development strategies of the western world. Knowledge becomes the most important asset and the basic drive behind many of the top world economies. On the other hand modern information and communication technologies are supplying us with more and more new and exciting ways of completing various everyday tasks. A large number of jobs that have for years been done in one way, thanks to the modern technology, are being done in a completely different, faster and simpler, way. The technology is affecting the teaching and learning process in the same way. Although there are many ways of implementing new technologies to aid the teacher, it is not altogether clear what is the best usage for faster, simpler and more efficient teaching [2], [3]. This means this topic is currently in focus of many researchers.

Most commonly used computer-aided teaching assessment methods today requires the teacher to compose the database of problems and questions that is then presented to the students in various forms. Problems usually take the form of multiple-choice questions (MCQ) that are easily inputed into the database, and are automatically graded [4]. Although there are many benefits to the technology, there is also much reason for caution. Firstly, providing such materials to the students without proper supervision and control is of doubtful use. The technology enables the student to access the material and solve the on-line quizzes in the environment that is **best suited based on their opinion**, thus seriously impairing productivity. Secondly, teachers and faculty, due to the many practical limitations such as limited time, low motivation etc., tend to create the contents once and the reuse it as often as possible. For assessment purposes this is clearly not desirable as the students than have the initiative to memorise, and not to develop full understanding of the taught matter [5].

Signals and Systems is a 4th semester base course taught at the University of Zagreb, Faculty of Electrical Engineering and Computing. During the years 2006/07 and 2007/08 we were involved in the effort to modernise the course and introduce new computer-aided methods that were deemed necessary due to the large increase in the number of students enrolled[1]. Main focus was on the assessment techniques. In 2006/07 numerous unsupervised on-line short quizzes were used, and in 2007/08 quizzes were administered under supervision. Multiple-choice questions (MCQ) were used both times. As the intended purpose of the quizzes was to provide feedback about learned matter (formative assessment) the overall score only contributed to the 5% of the final grade.

In this paper we compare the student results from the short multiple-choice questions quizzes with the student results on the classical written exam that we take as the baseline result. We show that there is no significant correlation between the MCQ quizzes and the classical written exam, regardless of the quiz delivery method, although supervised delivery improved the results. The paper is organised as follows: in the section II we give the overview of the different assessment methods, next section III discusses on how the course was organised. Section IV presents results that we discuss in section V.

II. Knowledge examination

The main task of the educational process is enabling students to learn skills and understand the material as set by the teacher. It is well understood that encouraging continuous work during the educational process improves knowledge retention [6].

As a measurement tool for the gained knowledge formative and summative assessments are used. Formative assessments provide feedback about the learners progress to both the student and the teacher, but with the grading aspect removed or minimized. Summative assessments are used to evaluate what has been learned and are used for formal grading [7]. To facilitate continuous work both formative and summative assessment must be continuous, meaning instead of one final examination as many small assessments as possible should be administered. Total number and methods of such small assessments have crucial role in determining the success of the teaching as they shape the learning process [4].

[1] The reformation of the curriculum caused the increase of the enrolled students from about 150 to about 800.

A wide selection of assessment methods is possible [1], [4]:

1) **Essays.** Used for testing understanding, synthesis and interpretation. Grading is time-consuming and objectiveness is difficult to achieve.
2) **Small projects.** Can evaluate all cognitive abilities. Objective grading of different projects is almost an impossible task.
3) **Oral presentations.** Used to evaluate communications skills.
4) **Questionnaires and check-lists.** Used as a guide to the student preforming a particular task.
5) **Reports.** Used to measure analysis and interpretation skills.
6) **Oral exams.** Test communications skills, require quick thinking on the student side. Provide immediate feedback. Can be subjective.
7) **Multiple-choice questions exams.** Can measure understanding, analysis and problem solving skills. Easy to administer. Questions should be precise, thus extending more time and effort on the teacher side during exam preparation.
8) **Short answer exams.** Can measure analysis, application and problem solving skills. Easier to prepare then MCQ questions, but more difficult to score.
9) **Problem-type exams.** Can measure application, analysis and problem solving strategies. Grading is difficult, especially achieving consistency.

Some of the above-mentioned methods are not designed or are not applicable for on-line computer-aided testing. For computer-aided testing multiple-choice questions (MCQ) are often used due to the simplicity and ease of delivery, scoring and feedback. Same applies for the written variants, especially if the standardised answer sheets and optical readers are used. When testing a large number of students there is almost no other viable alternative to the multiple-choice questions test [6].

Assessment methods are chosen to enable fair and objective testing and to provide accurate information what students are learning. For large number of students effort involved and time required for both teachers and students is also an important factor when selecting an assessment method [8].

Central role in assessing the knowledge are thorough **deep** summative examinations where not only analysis skills, but also synthesis skills must be shown[2]. For technical sciences such assessments are usually composed of more difficult short problems that are made **from scratch** for each new assessment.

In contrast to deep assessments there are also assessments designed to test **surface** learning where the students are required to demonstrate the knowledge recall and manipulation, and sometimes also application. Those were administered as short MCQ tests with the primary objective being encouragement of students to work continuously. To make the test more of a formative examination total weight in the final grade should be low.

A. Comparison of MCQ tests delivery methods

MCQ tests can be delivered in several different ways. Three are of particular interest: (1) first and the simplest one is computer-aided unsupervised delivery when students can solve the tests either alone or in groups. The main purpose of such assessments should be a formative one, providing feedback to both the student and the teacher. (2) Second delivery method assumes the students know what possible questions are, ie. by having access to the large database of questions. Short MCQ tests consisting of a small sample of the available questions are then delivered in controlled environment. (3) Third delivery method assumes the students know what material is to be tested, but the MCQ questions for the actual examinations will be made from scratch.

III. Organisation of the Signals and Systems course

Main objectives of the "Signals and Systems" course are: (a) to introduce the concepts of signals/systems, and (b) to apply those concepts in real-world situations in the fields of electrical engineering, telecommunications and computer science. The following outcomes are intended and help to identify knowledge gained by the student—after successfully completing the course the student will be able to

1) grasp the basics concepts of both continous and time-discrete signals,
2) be familiar with the required mathematical tools (Laplace and $\mathcal{Z}$-transforms), and
3) model and analyse linear systems in both time and frequency domain.

Although there are some examinations into alternative possibilities in teaching Signals and systems [10], [11], the main focus of any such course should be on assessment as the course forms a foundation for many different subfiles of electrical engineering.

In teaching Signals and systems following assessment methods were used: (1) computer aided formative multiple-choice questions assessments, (2) written summative multiple-choice question assessments, (3) written formative problem-type questions assessments, (4) written summative problem-type questions assessments, and (5) short oral assessments during laboratory sessions. Two midterms administered in a form of MCQ tests with newly designed questions and one final exam administered in a form of a problem-type exam are considered to be deep summative examinations. MCQ tests were also used for almost all formative examinations, with the only exception being oral examination during the laboratory sessions.

Such examination was predetermined by the total available time for the faculty. Due to the high number of the enrolled students MCQ exams were required as the classical written essays or problem type exams would be impractical to grade. From our experience the overall time required to carefully assemble the MCQ exam is significantly lower then the time required to carefully grade classical written exams.

[2] For the overview of the educational objectives taxonomy see Bloom, 1956 [9].

	2006/07	2007/08
Attendance	-	2
Active student participation	5	3
Short quizzes	5	5
Laboratory	10	10
Midterm 1	20	20
Midterm 2	20	20
Final exam	40	40

TABLE I

GRADING WEIGHTING SCHEME FOR VARIOUS COURSE COMPONENTS

	MCQ	1st midterm	2nd midterm	Final
MCQ	–	312.61	563.4	438.3
1st midterm		–	480.19	583.93
2nd midterm			–	656.59
Final				–

$\chi^2_{0.05}(81) = 103.01$

TABLE IV

H-VALUES FOR THE INDEPENDENCY TEST FOR 2006/07 EXAMINATIONS. CRITICAL VALUE IS 103.0 SO H_0 IS TO BE REJECTED FOR ALL EXAM COMBINATIONS.

	MCQ	1st midterm	2nd midterm	Final
MCQ	–	568.63	649.14	604.27
1st midterm		–	552.84	488.14
2nd midterm			–	618.59
Final				–

$\chi^2_{0.05}(81) = 103.01$

TABLE V

H-VALUES FOR THE INDEPENDENCY TEST FOR 2007/08 EXAMINATIONS. CRITICAL VALUE IS 103.0 SO H_0 IS TO BE REJECTED FOR ALL EXAM COMBINATIONS.

IV. RESULTS

Any comparison of examination results should be done under the assumptions that the results are indeed comparable. For a single subject comparisons of exam results is simpler as one can argue that the only important changing factor are the students themselves. So the comparison can be based on the assumption that for any single student and one only selected subject all exams taken by the student should have similar score. Such description would fit an ideal but never changing and never improving student, and is probably not applicable. However, if the course has a large number of students expected exam results should be similar. For real-world collected data it will be unfeasible to check all the underlying assumptions when comparing the exam results. Indeed, the comparisons are almost always made regardless whether they are justified or not.

One semester is relatively short period so we can assume that the average student score on the exams will not change much thus enabling comparison of the scores from various exam. For such comparison one would expect all exams scores to be dependant and, ideally, drawn from the same statistical distribution. To check the validity of those hypothesis we test: (a) are the resulting scores for any two exams statistically independent, and (b) are the resulting scores for any two exams drawn from the same distribution. As administered exams have different scoring scales to facilitate comparison all scores are linearly scaled to the range between 0% and 100%, with the 100% corresponding to the perfect score. Scores are divided into 10 categories, each spanning 10% of the total, so first category encompasses exam scores falling between 0% and 10% etc.

Table II shows obtained data for the comparison of the results on the MCQ test and the final examination for the independency test, while the table III shows obtained data for the same-distribution test. Real-world data collected on a single subject during the years 2006/07 and 2007/08 does not support those hypothesis—for all exam combinations both of the stated hypothesis are rejected. Results of analysis are summarised in the tables IV, V, VI and VII. The test statistics critical values for a significance level 5% are given below the table data.

Short quizzes were administered as MCQ exams, in both on-line and written form. During the year 2006/07 they were composed of computer-aided on-line formative weekly examinations followed by a summative one to form the grade. During the next year 2007/08 summative quizzes were administered as off-line written multiple-choice question exams with the formative preparatory exams being available in the same form as the year before. As can be seen in the table I short quizzes only contribute to the 5% of the total grade. The intention of the short quizzes is to provide immediate feedback about the factual knowledge to the student without creating extra pressure, so low contribution to the total grade is justified. During 2006/07 such short quizzes were available on-line both for practicing and for formal assessments.

When analysing the data from the computer-aided MCQ exams there are several variables of interest: 1) achieved score, and 2) time taken, and 3) number of repetitions (only for the formative MCQ exams). Of those time taken to complete the test is of particular interest as it is completely different for the formative and summative assessments. Figures 1 and 3 show the average time student spends in solving the formative exams[3]. During 2006/07 exams had 7 questions with the alloted time set to 15 minutes. Next year the number of questions was increased to ten. What is clearly a serious problem is the time students spend practicing, for 2006/07 average time is 1.03 minutes

[3] Exams are graded, but the grade is not used for formal evaluation.

	MCQ	1st midterm	2nd midterm	Final
MCQ	–	785.37	140.77	515.56
1st midterm		–	658.89	259.68
2nd midterm			–	274.08
Final				–

$\chi^2_{0.05}(18) = 9.3905$

TABLE VI

H-VALUES FOR THE SAME-DISTRIBUTION TEST FOR 2006/07 EXAMINATIONS. CRITICAL VALUE IS 9.3905 SO H_0 IS TO BE REJECTED FOR ALL EXAM COMBINATIONS.

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	$\sum$
b_1	22	0	0	0	0	0	0	0	0	0	22
b_2	27	0	0	0	0	0	0	0	0	0	27
b_3	32	1	1	0	0	1	0	0	0	0	35
b_4	33	1	0	0	0	0	0	0	0	0	34
b_5	38	4	2	3	0	0	2	1	0	0	50
b_6	33	5	2	3	4	1	0	0	0	0	48
b_7	42	11	6	11	6	12	5	2	0	0	95
b_8	48	11	16	15	14	16	13	9	3	0	145
b_9	32	13	12	12	19	30	31	24	16	0	189
b_{10}	11	1	9	12	20	16	26	33	14	7	149
$\sum$	318	47	48	56	63	76	77	69	33	7	794

$$H = 438.3 > \chi^2_{0.05}(81) = 103.01$$

TABLE II

ONLINE MCQ TEST AND FINAL EXAM FOR 2006/07. a_i AND b_i BOTH DENOTE i-TH BIN, IE. a_1 (OR b_1) ENCOMPASSES SCORES BETWEEN 0% AND 10%.

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	$\sum r_i$
X_1	22	27	33	36	47	51	90	143	196	149	794
X_2	307	48	50	56	63	69	84	63	45	9	794
$\sum$	329	75	83	92	110	120	174	206	241	158	1588

$$H = 515.56 > \chi^2_{0.05}(18) = 9.3905$$

TABLE III

ONLINE MCQ TEST AND FINAL EXAM FOR 2006/07. r_i DENOTES i-TH BIN, IE. r_1 ENCOMPASSES SCORES BETWEEN 0% AND 10%.

	MCQ	1st midterm	2nd midterm	Final
MCQ	–	442.01	203.77	154.61
1st midterm		–	127.07	366.7
2nd midterm			–	141.93
Final				–

$$\chi^2_{0.05}(18) = 9.3905$$

TABLE VII

H-VALUES FOR THE SAME-DISTRIBUTION TEST FOR 2007/08 EXAMINATIONS. CRITICAL VALUE IS 9.3905 SO H_0 IS TO BE REJECTED FOR ALL EXAM COMBINATIONS.

with the deviation of 1.34 minutes. **For whatever reason students do not like to practice!** In 2007/08 we stressed the importance of practice, but the average time increased marginally to 3.33 minutes with the deviation of 2.62 minutes. Again, this is clearly not enough, especially when compared to the summative MCQ exam average time of 9.99 minutes with the deviation of 4.15 minutes for 2006/07[4].

MCQ results can be compared to the other examination results. As MCQ are used as a formative assessment tool strong requirement for quality feedback would be relative similarity when compared against the summative results. For 2006/07 and 2007/08 scattergrams showing the overall **formative** on-line MCQ exams against the **summative** exams[5] are shown in figures 5 and 6. That there is no

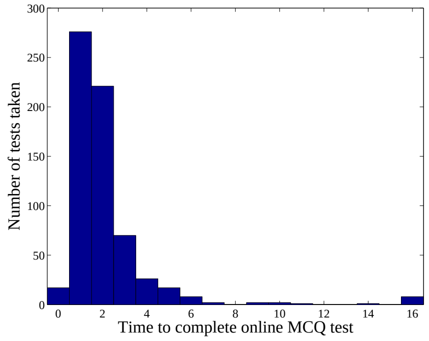

Fig. 1. Histogram showing on-line practice test taking times for 2006/07. The peak is at 1 minute meaning student takes about 9 seconds per question what is insufficient for training exam.

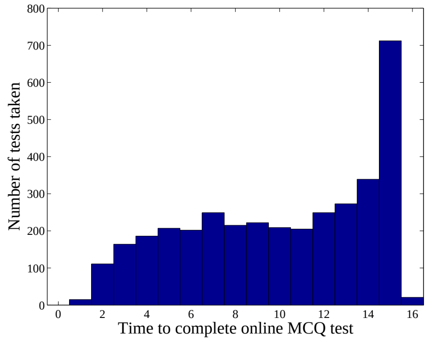

Fig. 2. Histogram showing on-line formal test taking times for 2006/07. The peak is now at the 15 minutes (maximal allowed time).

[4] For the 2007/08 exams were not administered on-line so the time take is not available.

[5] Two written MCQ exams for the midterms, written problem-type exam for the final.

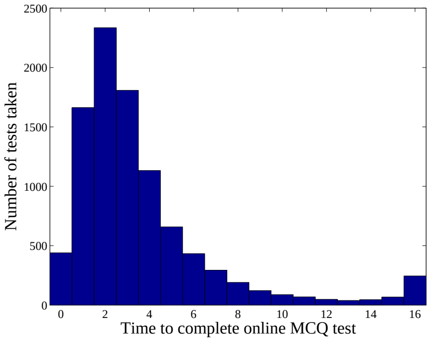

Fig. 3. Histogram showing on-line practice test taking times for 2007/08. The peak is at 2 minutes meaning student takes about 16 seconds per question what is insufficient for training exam.

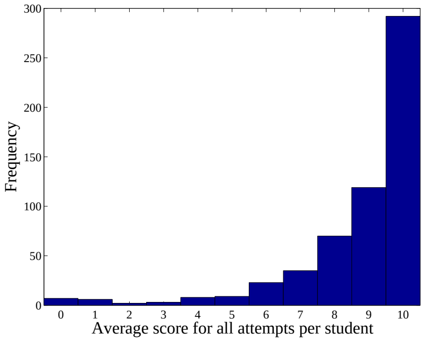

Fig. 4. Histogram showing achieved score on the practice quizzes for 2007/08 (same exam as in fig. 3). Distribution is skewed towards maximal score.

correlation is obvious, but even worse is the triangular shape the scattergram takes: **On-line formative exams systematically overestimate the knowledge the students posses!** The causes can be various are difficult to quantify.

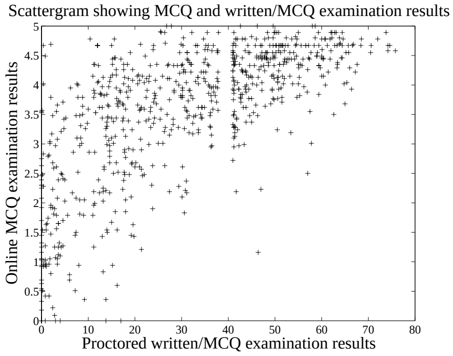

Fig. 5. Scattergram for the 2006/07 exam data showing on-line administered non-proctored MCQ exams vs. proctored written examination results. The data is grouped above the diagonal indicating a probable bias for the on-line MCQ examinations.

For the further comparison boxplots can be used[6]. For the summative and on-line formative exams we would like to have a match between average values and the deviations, with the as good overlap as possible. For the 2006/07 the 1st midterm and final exam are consistent in that way, with the formative on-line tests and the 2nd midterm overstating the student achievement (figure 8). For the 2nd MCQ midterm students were given similar MCQ questions in advance, and that is probable cause for the average overestimate[7]. For the 2007/08 the results are much more consistent, except the first formative assessment, due to the several reasons: 1) formative MCQ were administered in controlled environment, and 2) students could prepare better due to the course material prepared in the 2006/07 that was made available on-line.

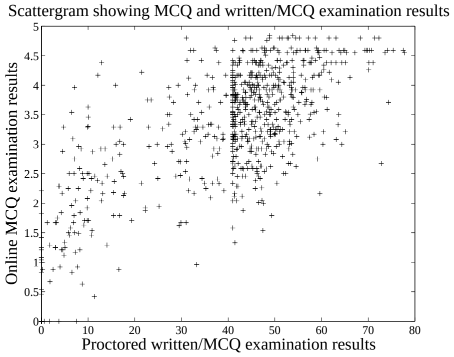

Fig. 6. Scattergram for the 2007/08 exam data showing on-line administered non-proctored MCQ exams vs. proctored written examination results. The data is grouped above the diagonal indicating a probable bias for the on-line MCQ examinations.

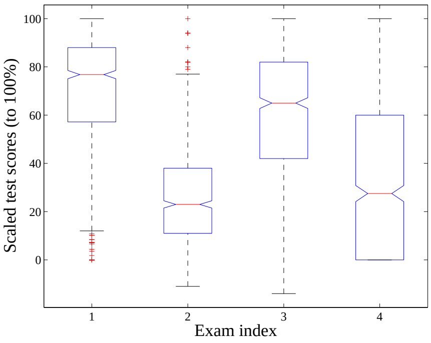

Fig. 8. Boxplot showing examination scores for 2006/07. Index 1 are short quizzes, indices 2 and 3 and 1st and 2nd midterm and 4 is final exam.

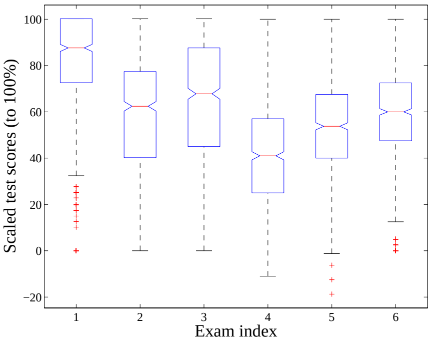

Fig. 9. Boxplot showing examination scores for 2007/08. Indices 1-3 are short quizzes, 4 and 5 and 1st and 2nd midterm and 6 is final exam.

[6] ANOVA or similar analysis tool would be better for analysis, but the required assumptions of normality, independence and equality are not met by the collected real-world data.

[7] Students memorised the required procedures to solve the practice problems without proper understanding.

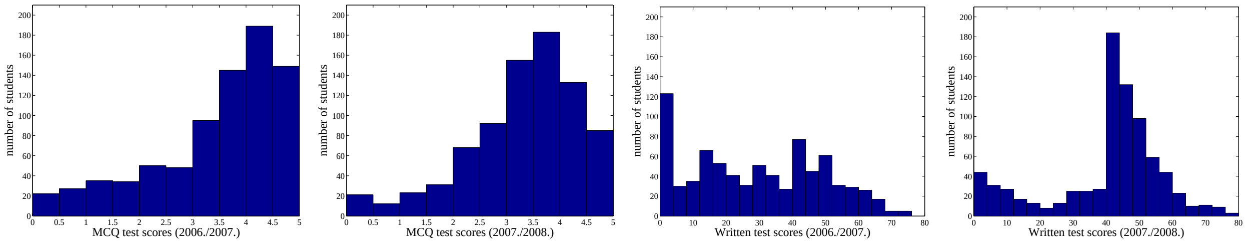

Fig. 7. Histograms showing the overall score distribution for different test delivery methods

V. Discussion

As demonstrated by the results first delivery method (see section II-A) in real world application has doubtful value, both in formative and summative assessments. Even awarding a small number of points to encourage the students has devastating effect as the students focus not on learning, but on achieving the highest possible score with the minimal effort.

Second delivery method is better, but still lacking if accurate measure of the student knowledge is desired. Students are now more inclined when trying understand the matter, but can also tend to only memorise more difficult questions thus achieving the balance between effort and obtained score. If the MCQ are delivered in such way that for each examinations the questions are made from scratch (difficult and time consuming) students can not benefit if they memorise the questions, and must instead extend more effort to obtain proper understanding. If MCQ test are used as formative assessments should the result affect the final mark? As the final goal of the student is to pass assigning a low weight to the formative assessments has a negative effect in a sense that students are not interested as they do not perceive any **immediate** gain. Assigning too high a weight has opposite effect as it increases the student interest, but also encourages all other behaviours that can yield higher score.

VI. Conclusion

Obtained data shows that there is little value in administering both formative and summative assessment in a form of MCQ exams without proper supervision, especially if the number of enrolled students in a course is large. Formative value of the unsupervised MCQ on-line tests is also doubtful, especially when considering the average time the students actually spend on assessment was shown to be incredibly short. If we want the MCQ exams to measure the knowledge more accurately several conditions must be met: (1) both formative and summative MCQ exams must be administered in controlled environment, (2) student must perceive either immediate gain or immediate loss in the total grade depending on the results of the exam, and (3) for each administered MCQ test questions should be constructed from scratch without reusing existing question databases. Although the third requirement places additional burden on the teachers, in our experience time required is still less than the time required for the grading of the classical written exams.

Presented work is not intended to present a critique of the technology, but to be a strong reminder that the technology should be carefully used!

VII. Acknowledgement

We would like to extend our gratitude to the professors Branko Jeren and Damir Seršić for their guidance, support and many usefull discussions and suggestions that made this work possible.

References

[1] K. P. Cross and T. A. Angelo, *Classroom Assessment Techniques. A Handbook for Faculty*. National Center for Research to Improve Postsecondary Teaching and Learning, The University Of Michigan, Sept. 1988.

[2] R. McGreal and M. Elliott, *Theory and Practice of Online Learning*. Athabasca University, 2004, ch. Technologies for Online Learning (E-Learning). [Online]. Available: http://cde.athabascau.ca/online_book/ch5.html

[3] A. Bork and D. R. Britton Jr., "The web is not yet suitable for learning," *IEEE Computer*, vol. 31, no. 6, pp. 115–116, June 1998.

[4] G. Brown, "Assessment: A guide for lectures," LTSN Generic Center (Learning and Teaching Support Network), York, Tech. Rep., Nov. 2001. [Online]. Available: http://www.heacademy.ac.uk/resources/detail/ourwork/tla/assessment_series

[5] S. Brown, P. Race, and B. Smith, *500 Tips on Assessment*, 2nd ed. New York: RoutledgeFalmer, 2005.

[6] J. Biggs, *Teaching for quality learning at university: what the student does*, 2nd ed. SRHE and Open University Press, Feb. 2003.

[7] P. Knight, "A briefing on key concepts: Formative and summative, criterion and norm-referenced assessment," LTSN Generic Center (Learning and Teaching Support Network), York, Tech. Rep., Nov. 2001. [Online]. Available: http://www.heacademy.ac.uk/resources/detail/ourwork/tla/assessment_series

[8] G. Gibbs, *Assessing more students*. Polytechnics and Colleges Funding Council, 1992.

[9] B. S. Bloom, Ed., *Taxonomy of Educational Objectives: The Classification of Educational Goals by a Committee of College and University Examiners*. New York: McKay, 1956.

[10] V. Stonick, "Teaching signals and systems using the virtual laboratory environment in ece at cmu," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1, pp. 36–39, 1993.

[11] R. F. Vaz and N. Arcolano, "Teaching signals and systems through portfolios, writing, and independent learning," in *Proceedings of ASEE 2001*, Albuquerque, New Mexico, June 2001.