Comparison of Electronic Examinations using Adaptive Multiple-choice Questions and Constructed-response Questions

Peter J. Stavroulakis[1,2,3] [a], Panagiotis Photopoulos[1] [b], Errikos Ventouras[4] [c] and Dimos Triantis[1] [d]

[1]Department of Electrical and Electronic Engineering, University of West Attica, Athens, Greece
[2]Department of Management and International Business, School of Business and Economics,
The American College of Greece, Ag. Paraskevi, Greece
[3]Department of Maritime Studies, School of Maritime & Industrial Studies, University of Piraeus, Piraeus, Greece
[4]Department of Biomedical Engineering, University of West Attica, Athens, Greece

Keywords: Evaluation Methodologies, Adaptive Testing, Multiple-choice Questions, Electronic Examinations.

Abstract: The use of computer-based examination systems offers advantages related to the reduction of human resource allocation and to gains in objectivity for the scoring process. Multiple-choice questions (MCQs) are widely used in such systems; one of the main reasons that hamper their effectiveness, in comparison with constructed response questions (CRQ), is the influence of guessing. Considering limitations within previously proposed MCQs examination methods and scoring rules, in the present work a novel MCQs examination method is presented, termed 'adaptive' MCQs method. MCQs are divided into 3 categories, per difficulty level. The 'path' that an examinee will follow is constituted by 3 phases, wherein a set of questions belonging to one of the three difficulty-categories, is appointed. The exact path followed is selected per the success level of the examinee in the preceding phase. The scoring provided by the adaptive MCQs examination method produced results that were statistically indistinguishable to the scoring produced by a traditional CRQ examination method. At the same time, both the scoring results of the adaptive MCQs examination and the scoring results of the CRQ examination differed significantly from those obtained by a generic 'non-adaptive' MCQs examination.

1 INTRODUCTION

New techniques, approaches and technologies are constantly introduced within the educational framework (McMorran, Ragupathi, and Luo 2017), to the point that this phenomenon might constitute a paradigm shift, at least with respect to educational procedures (Nulty 2008). The introduction and use of technology within educational procedures requires prudent effectiveness assessment (Ćukušić, Garača, and Jadrić 2014). Empirical evidence points towards effectiveness of methods (Howard, Schenk, and Discenza 2004). Among other topics, the discourse is especially concerned with assessment and examination methods, to provide indications of pertinent instruments (Desrochers and Shelnutt 2012), as well as investigation of methodologies for problem solving with novel assessment tools (Adesina et al. 2014).

Adaptivity has been actively pursued in the context of computer-based education and testing. In the context of testing, computerized adaptive testing (CAT) provides alternative paths of action to the examinee. This may be attained by calculating an estimate of the examinee's competence, after each answer, and then providing the next question according to this estimate (Thissen and Mislevy 2000). CAT methodologies have undergone significant evolution with many promising results (Van der Linden and Glas 2000), for example CAT systems based on the Item Response Theory (IRT) (Hirose et al. 2016) and on Bayesian Networks (Culbertson 2016).

[a] https://orcid.org/0000-0003-4545-3102
[b] https://orcid.org/0000-0001-7944-666X
[c] https://orcid.org/0000-0001-9826-5952
[d] https://orcid.org/0000-0003-4219-8687

Stavroulakis, P., Photopoulos, P., Ventouras, E. and Triantis, D.
Comparison of Electronic Examinations using Adaptive Multiple-choice Questions and Constructed-response Questions.
DOI: 10.5220/0009341603580365
In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020) - Volume 1*, pages 358-365
ISBN: 978-989-758-417-6

Within educational frameworks, examples of applications of adaptive testing include language placement examinations (Stahl, Bergstrom, and Gershon 2000), dynamic assessment through personalized two-tier test questions that help teachers discover the reasoning of students (Liu et al. 2007), assessment of programming abilities (Syang and Dale 1993), and CAT using out-of-level-testing, that improved measurement accuracy and test efficiency for students who perform significantly above or below their grade-level peers (Wei and Lin 2015).

One of the main incentives for using electronic examinations is related to the possibility for automatic scoring, when using true/false statements and multiple-choice questions (MCQs) (Scharf and Baldwin 2007). The most widely used MCQs variant belongs to the 'single-best answer' scheme, in which the examinee must select the most appropriate single response per question, from a set of responses/answers (McCoubrie and McKnight 2008). MCQs' advantages have been extensively investigated (De-Marcos et al. 2010; Scharf and Baldwin 2007; Van der Linden and Glas 2000).

Guessing in some answers in MCQs will inevitably result in collecting partial scores in the final score, without possessing knowledge of the questioned material. As a remedy for this situation, various alternative marking schemes have been proposed, including 'mixed-scoring' methods and others, wherein students collect points for correct answers and lose points for incorrect answers (Scharf and Baldwin 2007).

The evaluation of various examination methods focuses on the quality of their results, quality being expressed in terms of reliability and validity. Reliability might be defined as the degree to which a test repeatedly yields grading marks that truly reflect the understanding and knowledge of the examinee (Downing 2004). Validity might be defined as the ability of the testing method to measure the educational outcome that it was intended to test (Downing 2003). Reliability and validity indicators have been actively investigated for examination methods (Wass et al. 2003).

Research has produced substantial indications that MCQs tests provide grades that have comparable validity to grades produced by CRQ examinations, and might even possess higher reliability than CRQ tests (Lukhele, Thissen, and Wainer 1994; Wainer and Thissen 1993). Although it is accepted that MCQs have a predominant positive constituent of objectivity, there are cases where their validity may be questioned, such as in the case of the clearly quantifiable grade bias introduced by the 'positive-grades-only' scoring rule, or in the more subtly interfering processes related to the 'mixed-scoring' rules. Nevertheless, because MCQs greatly contribute in reaching objectivity and standardization of the examination process, they are often preferred over more traditional examination methods, such as CRQ examinations.

To alleviate the problems related to MCQs examinations using 'positive-grades-only' or 'mixed-scoring' rules, a novel kind of MCQs examination method has been proposed, which uses sets of pairs of MCQs, referred as 'paired' MCQs method (p-MCQs) (Triantis and Ventouras 2012; Ventouras et al. 2011). The questions' answers are graded in pairs, providing a bonus (or penalty) if both (or only one) answer(s) of the pair are answered correctly. The above research provided indications regarding the ability of the p-MCQs method to surpass the limitations of the simplest grading scheme of MCQs, while at the same time avoiding the 'direct' negative markings of multiple-choice items and their concomitant negative effects. Nevertheless, a rather demanding requirement is posed on the construction of the MCQs data bank, namely that for each examined topic several pairs of questions of equivalent level of difficulty should be constructed, in such a way that the fact that each MCQ of the pair concerns the same topic should not be evident to a student who does not possess adequate knowledge on the topic addressed in the questions of the pair.

Considering the limitations of previously proposed MCQs examination methods and scoring rules, in the present work a novel MCQs examination method is presented, termed 'adaptive' MCQs method. Its aim is to avoid the various explicit or implicit negative marking rules altogether. At the same time, while no negative marks are given for wrongly answered or omitted questions, the methodology of the adaptive MCQs method tries to circumvent the positive-grades bias of the 'positive-grades-only' rule. The main characteristics of the method are that it divides the examination material into three categories of MCQs, per difficulty level of the questions, reflected also in the weight that those questions will have in the scoring process, and the 'path' that an examinee will follow is established by 3 phases (pertaining to a 'three-tiered approach'), in which she/he is given a set of questions belonging to one of the 3 difficulty categories. The exact path that an examinee will follow, concerning the level of difficulty of MCQs in the 2nd and 3rd phase is not determined a priori, but is decided by the system, by adapting to the success level of the examinee in the preceding phase. The less well-prepared students, who are expected to be more prone to guessing, will remain in the low-difficulty low-weight set of MCQs. On the other hand, more well-prepared students will have the chance to pass to MCQs of increasing difficulty. Since those students are not expected to rely on guessing, their probability of achieving the higher scores related to the high-difficulty MCQs set is supposed to be augmented.

Therefore, an investigation was carried out concerning whether the grades given using the adaptive MCQs examination method are statistically indistinguishable to the grades given using the CRQs examination method, while differing from a simpler 'non-adaptive' MCQs examination. Both MCQs examination methods used the 'positive-grades-only' scoring rule. The three examinations were given to the same sample of students, on the same topics and with the same levels of difficulty.

2 MATERIALS AND METHODS

2.1 The Examined Course and the Sample of Students

Three examination methods were used, described in the sections that follow. The course in which the students were examined was an introductory course entitled 'Physics of semiconductor devices', which belongs to the group of core background courses and is taught at the Electrical and Electronics Engineering Department of the University of West Attica. A group of 46 students participated in the study, taking the 3 examinations.

All students had completed the course and were familiarized with the electronic examination platform used. All tests were administered from the same personnel and attention was given to mitigate issues that might constitute validity threats related to any attempts of external communication or cheating among the examinees. The examination took place in a PC laboratory room using the 'e-examination' application, a software package that utilizes CAT methodology developed at the University of West Attica (Tsiakas et al. 2007).

2.2 Examination Methods

At first, the students underwent a conventional CRQ examination in which they were given six questions to answer, that covered 80% of the syllabus taught throughout the semester; they were asked to select and answer four out of six questions. Students typed their answers into the appropriate text field. After the end of the pre-determined examination duration time, the answers were automatically formatted into pdf format files and were e-mailed to the examiner, and printed. Each examinee got a copy of her/his answers. The scores of the CRQ examination are denoted in the sections that follow by 'scrq'. These scores constituted the final students' marks for the course. The highest mark that could have been achieved was 100.

At a later stage, the same students were given two electronically administered MCQs tests. The MCQs were selected from within a MCQs database in a random manner, avoiding repetition of question items. A restriction was incorporated in the selection of questions, namely that the questions should cover at least 80% of the syllabus.

The first MCQs test ('MCQ-1') included 12 questions (denoted 'A') with weighting factor 1 (W=1), and 12 questions with greater difficulty (denoted 'B') and a weighting factor of 2 (W=2). The students were free to answer any of the questions, irrespective of their difficulty level or their performance level in prior questions. This characteristic of the examination constituted its 'non-adaptive' nature, in contrast to the adaptive examination scheme described in the next section. The scoring results from this test are denoted by 'smcq-1'. The highest mark that could have been achieved was 36 ($12\cdot1 + 12\cdot2 = 36$). For the analysis used in the present work, the marks of examination MCQ-1 were normalized to a maximum of 100. No negative marks are given for wrong answers or omitted questions.

The second MCQs test ('MCQ-2') consisted of a combination of possible examination paths comprising 3 examination phases. At each phase a set of MCQs was given. The sets of MCQs belonged to 3 types, per their level of difficulty, and were denoted by 'A', 'B' and 'C', in an ascending level of difficulty, respectively. The correct answer quota (CAQ) for passing/failing a phase was 0.5, i.e., 'pass' (or 'fail') meant to answer correctly at least (or less than) 50% of the questions given.

Students were not aware of their score at each phase and the passage from one set of questions to the next was done automatically, using the examination software. At the 1st phase, the students were asked to answer 12 type 'A' MCQs. In the 2nd and 3rd phase 12 and 3 MCQs were given, respectively, but the type of questions given (i.e., whether the questions given were of type 'A', 'B', or 'C') was adapted per the performance of the examinee in the 1st and the 2nd phase.

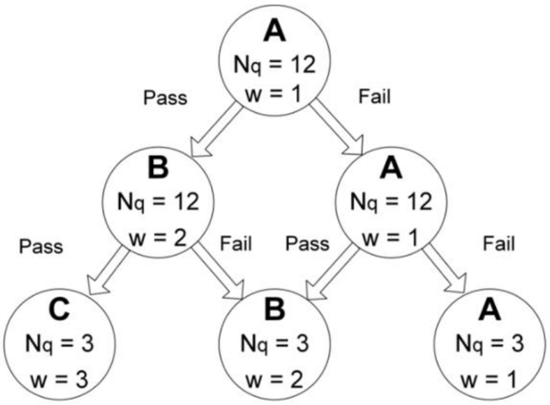

Figure 1: The possible examination paths of examination MCQ-2 and the rules followed for the transition from one phase to another. Nq is the number of questions in the set presented at each phase. W is the weighting factor of each correctly answered question, in the respective phase.

The examination paths that could be followed, after the 1st phase, are the following (graphically portrayed in Figure 1):

(1) Examination path leading from type 'A' MCQs to type 'B' MCQs and then to type 'C' MCQs, denoted in the following as path 'ABC': If the students answered correctly at least 6 of the type 'A' questions of the 1st phase (i.e., they reached a 'pass' status for the 'A' questions' set), then, at the 2nd phase, they were given a set of 12 type 'B' MCQs to answer. If the students answered correctly at least 6 of the type 'B' MCQs (i.e., they reached a 'pass' status for the 'B' questions' set), then, at the 3rd phase, they were given a set of 3 type 'C' MCQs.

(2) Examination path leading from type 'A' MCQs to type 'B' MCQs and then to type 'B' MCQs, denoted as path 'ABB': If the students answered correctly at least 6 of the type 'A' questions of the 1st phase ('pass'), then, at the 2nd phase they were given a set of 12 type 'B' MCQs to answer. If the students did not answer correctly at least 6 of the type 'B' MCQs (i.e., their performance is a 'fail' for the 'B' questions' set), then, at the 3rd phase, they were given another set of 3 type 'B' MCQs.

(3) Examination path leading from type 'A' MCQs to type 'A' MCQs and then to type 'B' MCQs, denoted as path 'AAB': If the students did not answer correctly at least 6 of the type 'A' questions of the 1st phase ('fail'), then, at the 2nd phase, they were given a set of 12 type 'A' MCQs to answer. If the students answered correctly at least 6 of the type 'A' MCQs ('pass'), then, at the 3rd phase, they were given a set of 3 type 'B' MCQs.

(4) Examination path leading from type 'A' MCQs to type 'A' MCQs and then to type 'A' MCQs, denoted in the following as path 'AAA': If the students did not answer correctly at least 6 of the type 'A' questions of the 1st phase ('fail'), then, at the 2nd phase, they were given a set of 12 type 'A' MCQs to answer. If the students did not answer correctly at least 6 of the type 'A' MCQs ('fail'), then, at the 3rd phase, they were given another set of 3 type 'A' MCQs.

The difficulty of the questions was based on the level of effort needed for the students to answer the questions, relating to the depth of knowledge they should possess in order to answer correctly, according to the course design principles used by the teachers who developed the course material. Additionally, the level of difficulty was ascertained, for the specific course, by the rates that students got to comparable questions in examination sessions of previous semesters, in accordance to their overall course performance.

The fact that the path that a student would follow is not determined prior to the start of the examination and is dependent on her/his performance in each of the two initial phases constituted the 'adaptive' nature of the examination method.

The score received is denoted by 'smcq-2' and is computed as follows for the different cases:

(1) Case ABC: smcq-2= N+ 2K + 3M, where N, K and M are the number of correct answers given by the student in the 1st, 2nd and 3rd phase, respectively. The range of N is 6 to 12, of K is 6 to 12 and of M is 0 to 3.

(2) Case ABB: smcq-2= N + 2K + 2M. The range of N is 6 to 12, of K is 0 to 5 and of M is 0 to 3.

(3) Case AAB: smcq-2= N + K + 2M. The range of N is 0 to 5, of K is 6 to 12 and of M is 0 to 3.

(4) Case AAA smcq-2= N + K + M. The range of N is 0 to 5, of K is 0 to 5 and of M is 0 to 3.

The highest mark that could have been achieved was 45, when the student's path belonged to case ABC and she/he answered correctly all questions (12·1+12·2+3·3=45). No negative marks were given for wrong answers or omitted questions. For the analysis used in the present work the marks of examination MCQ-2 were normalized to a maximum of 100.

2.3 Statistical Hypotheses

The comparison of the MCQ-2 examination method to the MCQ-1 examination method and to the CRQ method, which might be assumed to be the 'gold standard' method, aimed at providing indications for accepting the MCQ-2 examination method as an alternative for CRQ methods, according to the rationale stated in the Introduction section. Such an indication might be provided, if the scores obtained through the MCQs, using the MCQ-2 method, are statistically indistinguishable from the scores obtained from the CRQ method. In this line of thought, the null hypothesis (H0) to be tested in comparing the CRQ, the MCQ-1 and the MCQ-2 examination methods could be stated as: ‘The means of the distributions of scores scrq, smcq-1 and smcq-2 are equal’.

If hypothesis H0 is rejected, i.e., the overall differences between the three means are significant, then post-hoc pair-wise comparisons, with adjustment for multiple comparisons, should be used, in order to check the three ‘secondary’ hypotheses, namely H0 (CRQ to MCQ-1) (i.e., ‘The means of the distributions of scores scrq and smcq-1 are equal’), H0 (CRQ to MCQ-2) (i.e., ‘The means of the distributions of scores scrq and smcq-2 are equal’) and H0 (MCQ-1 to MCQ-2) (i.e., ‘The means of the distributions of scores smcq-1 and smcq-2 are equal’).

3 RESULTS

Table 1 presents descriptive statistical values concerning the examination methods used. For the MCQ-2 examination method, from the 46 students, 18 followed path ABC 19 followed path ABB, 0 followed path AAB and 9 followed path AAA.

Table 1: Descriptive statistics for the scores of the 3 examination methods.

	scrq	smcq-1	smcq-2
Mean (m)	50.63	60.14	51.03
Standard Deviation (S.D.)	19.98	17.04	20.00
Maximum value	90.00	94.44	91.67
Minimum value	15.00	30.56	16.67

The Kolmogorov-Smirnov goodness-of-fit test (p=0.20 for scrq, mcq-1 and mcq-2) and the Shapiro-Wilk test (p=0.14 for scrq, p=0.23 for mcq-1 and p=0.23 for mcq-2) showed that the distributions of the scores scrq, smcq-1 and smcq-2 were consistent with a normal distribution. Therefore, we might consider that all three sets of scores can be regarded as originating from a normal distribution. The histogram of the distribution of each scoring variable is shown in Figure 2.

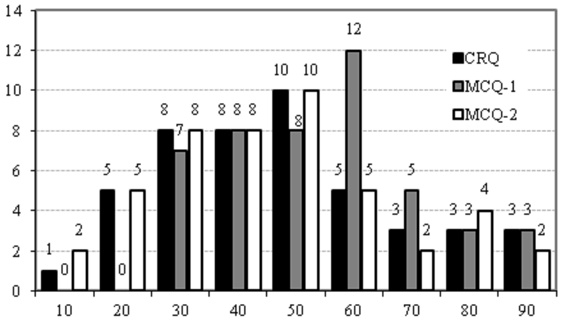

Figure 2: Number of students graded in the respective score ranges, per examination method. Black bars: CRQ examination method. White bars: MCQ-2 examination method. Grey bars: MCQ-1 examination method. Each score range extended from x to x+9.99.

Cronbach’s alpha coefficient was used as an internal consistency estimate of the reliability of the examination method scores. A high value for Cronbach’s alpha (α=0.994) and standardized Cronbach's alpha ($\alpha_{standardized}$=0.996) was observed.

For the whole set of students who undertook the examination, repeated-measures ANOVA with one within-subjects factor (method of examination, three levels) indicated that the within-subject effect was significant (F1.282,57.668=204.042, $p<0.001$, partial η^2=.819, degrees of freedom were corrected for non-sphericity according to the Greenhouse-Geisser procedure). ANOVA was followed by planned comparisons between each of the examination methods, assessed with post-hoc Bonferroni pair-wise comparisons at the 0.05 level of significance. Significant differences existed between scrq and mcq-1 ($p<0.001$) and between mcq-1 and mcq-2 ($p<0.001$). In addition to the above results, regression analysis using the Pearson coefficient of correlation between variables scrq, mcq-1 and mcq-2 indicated that adjusted R^2 (scrq, mcq-2) = 0.99 was greater than adjusted R^2 (scrq, mcq-1) = 0.97.

These results indicate that the MCQ-2 examination method (resulting in score smcq-2) is statistically equivalent to the CRQ examination method (resulting in score scrq). Both methods differ significantly from the MCQs examination method that does not use the ‘adaptive’ scheme for test item presentation, i.e., the MCQ-1 method (resulting in score smcq-1). The results also indicate clearly that students achieve greater scores with MCQ-1 compared to the scores achieved using both the CRQ examination method and the MCQ-2 examination method. This bias can also be deduced from the regression line of the score of MCQ-1 to the score of CRQ presented in Figure 3.

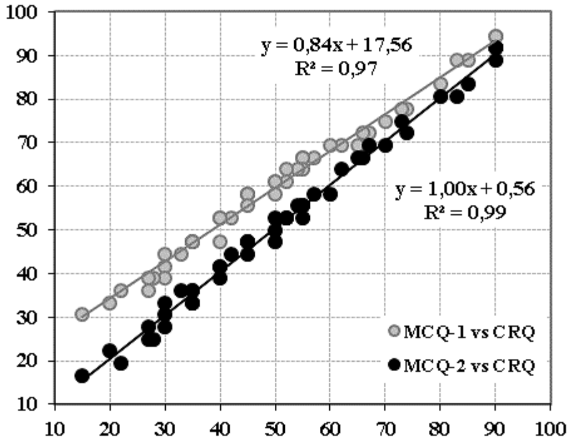

Figure 3: Regression line (y=0.84x+17.56) of score of MCQ-1 (smcq-1) to score of CRQ (scrq) and regression line (y=1.00x+0.56) of score of MCQ-2 (smcq-2) to score of CRQ (scrq).

A reason for this characteristic of the MCQ-1 method might be that MCQ-1 does not use scoring mechanisms (such as the one included in MCQ-2) designed to mitigate the effects of guessing on choosing correct answers. Similar results were obtained in previous research, concerning the comparison of MCQs to CRQ examination methods and Structured Oral examination methods (Ventouras et al. 2011). The adaptive scheme used for item presentation in the MCQ-2 examination method might be the reason for removing the bias present in MCQ-1. This is also indicated if the regression line of MCQ-1's score to CRQ's score is compared with the regression line of MCQ-2's score to CRQ's score (Figure 3).

4 CONCLUSIONS

The results presented in this study indicate that the examination method using the novel adaptive MCQs scheme can be effectively used in the framework of electronic examinations. The principles used in the design of the novel MCQs examination method were the following:

(1) The method should be constructed in such a way that the positive grade bias known to be introduced by MCQs grading methods using a positive-grades-only scoring rule is avoided.

(2) The hampering effects related to the introduction of negative marking schemes should also be avoided, because of the variance that might be introduced in the test scores related to the expectations of the examinees and not to the knowledge that is tested.

(3) The requirements for the construction of the question bank for the MCQ test should not become a primary dissuading factor for introducing the proposed examination method into everyday educational practice.

Indications are also provided that the requirements imposed on the adaptive MCQs method were successfully met. The proposed adaptive MCQs examination method (MCQ-2) provided scoring results that were statistically indistinguishable from the scores provided by the CRQ method. In contrast, the MCQs examination method using the positive-grades-only scoring rule in its most basic form (MCQ-1) produced significantly higher grades than the adaptive MCQs and the CRQ method.

Therefore, indications have been provided, in the framework of the present study, that the positive grade bias known to be introduced by MCQs grading methods using a positive-grades-only scoring rule is avoided. This is remarkable, since the proposed method also uses a positive-grades-only scoring rule. The use of a positive-grades-only scoring rule was deliberately included in the design phase of the study, since it was surmised that an effective way to avoid the distorting effects of negative marking schemes was to avoid negative marking completely, excluding even the indirect negative marking used in the p-MCQs method proposed in earlier research.

An explanation for the fact that the positive grade bias related to positive-grades-only scoring rule is avoided using MCQ-2, despite the fact that MCQ-2 uses such a rule, might be that, as stated also in the Introduction section, through the three-tiered scheme, the less well-prepared students, expected to be more prone to guessing, will probably be 'kept' in the low-difficulty low-weight set of MCQs, while the more well-prepared students will probably pass to MCQs of gradually increasing difficulty and scoring weight. Therefore, a grater score was expected to reflect more faithfully an actually higher level of knowledge of the examinee, compared to what would happen if no adaptive scheme was used.

The introduction of MCQs in the educational procedures is accompanied by a need for an initial allocation of a significant amount of resources of experienced teachers, for constructing a suitable question data bank, irrespectively of the MCQs scheme used. This will concern an adequately large coverage of the range of the examined topic, with non-overlapping questions and with a difference between the selection choices that will be sufficiently clear to a well-prepared student. Therefore, when MCQs examinations are planned to be introduced, a compromise should be pursued between, on the one hand the ease of producing the MCQs and, on the other hand, the specific requirements and variants that the MCQs examination method presents. In this context, the only requirement concerning the construction of the question bank which goes beyond the 'basic' compilation requirements, for the adaptive MCQs, is that 3 sets of questions should be constructed, clearly differing with respect to the level of difficulty of the questions they contain.

The statistical coincidence of the adaptive MCQs examination method scores with the CRQ examination method scores provides a degree of assurance about the soundness of the choices that were made in designing the adaptive method, concerning the number of questions included in each of the 3 phases of the examination procedure, as well as the range of the weighting factors. Nevertheless, future investigations, in addition to extending the application to larger sets of students and to other topics, should include variations in the previously mentioned parameters, especially to check in a more in-depth way the effects of the inclusion of type C questions in the 3rd phase of the examination procedure. The attention paid to these types of questions assumes that their presence is the most effective way to alleviate the positive-grade bias related to the positive-grades-only rule used. The optimum compromise should also be investigated, between the differences in the level of difficulty of the 3 types of questions so as, on the one hand to restrict less-well-prepared students from acquiring partial scores due to guessing and, on the other hand, to avoid dissuading effects due to the augmentation of the difficulty in type B questions and, perhaps more importantly, in type C questions, that might occur even to well-prepared students.

In conclusion, keeping in mind the restrictions of the present investigation concerning the number of students that participated in the study, as well as the fact that it was applied to a specific course, the present work provides indications that the principles used for designing the adaptive MCQs examination method achieved their aim to a satisfactory degree, since the positive-grades bias was avoided, no negative marking scheme was used, neither explicitly nor implicitly, and the requirements for constructing the question bank were not substantially augmented, in comparison to the standard requirements imposed by the implementation of any generic MCQs examination method.

REFERENCES

Adesina, A., R. Stone, F. Batmaz, and I. Jones. 2014. "Touch Arithmetic, A process-based Computer-Aided Assessment approach for capture of problem solving steps in the context of elementary mathematics." *Computers and Education* 78, 333-343.

Ćukušić, M., Z. Garača, and M. Jadrić. 2014. "Online self-assessment and students' success in higher education institutions." *Computers and Education* 72, 100–109.

Culbertson, M. J. 2016. "Bayesian networks in educational assessment, The state of the field." *Applied psychological measurement* 40, 3-21.

De-Marcos, L., J. R. Hilera, R. Barchino, L. Jiménez, J. J. Martínez, J. A. Gutiérrez, S. Otón, et al. 2010. "An experiment for improving students' performance in secondary and tertiary education by means of m-learning auto-assessment." *Computers and Education* 55, 1069-1079.

Desrochers, M. N., and J. M. Shelnutt. 2012. "Effect of answer format and review method on college students' learning." *Computers and Education* 59, 946-951.

Downing, S. M. 2003. "Validity, On meaningful interpretation of assessment data." *Medical Education* 37, 830-837.

Downing, S. M. 2004. "Reliability, On the reproducibility of assessment data." *Medical Education* 38, 1006-1012.

Hirose, H., Masanori T., Yusuke Y., Tetsuji T., Tatsuhiro H., Fujio K., Mitsunori I., and K. Tetsuya. 2016. *Questions and Answers Database Construction for Adaptive Online IRT Testing Systems, Analysis Course and Linear Algebra Course. Proceedings of the 5th IIAI International Congress on Advanced Applied Informatics*, 433-438. Piscataway, New Jersey, IEEE.

Howard, C., K. Schenk, and R. Discenza. 2004. *Distance Learning and University Effectiveness, Changing Educational Paradigms for Online learning*. London, Information Science Publishing.

Liu, Y. C., M. C. Chiang, S. C. Chen, V. Istanda, and T. H. Huang. 2007. *An online system using dynamic assessment and adaptive material. Proceedings of the 37th Annual ASEE/IEEE Frontiers in Education Conference*, T3D6-T3D10. Piscataway, New Jersey, IEEE.

Lukhele, R., D. Thissen, and H. Wainer. 1994. "On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests." *Journal of Educational Measurement* 31, 234-250.

McCoubrie, P., and L. McKnight. 2008. "Single best answer MCQs, a new format for the FRCR part 2a exam." *Clinical Radiology* 63, 506-510.

McMorran, C., K. Ragupathi, and S. Luo. 2017. "Assessment and learning without grades? Motivations and concerns with implementing gradeless learning in higher education." *Assessment and Evaluation in Higher Education* 42, 361-377.

Nulty, D. D. 2008. "The adequacy of response rates to online and paper surveys, What can be done?"

Assessment and Evaluation in Higher Education 33, 301-314.

Scharf, E. M., and L. P. Baldwin. 2007. “Assessing multiple choice question (MCQ) tests - a mathematical perspective.” *Active Learning in Higher Education* 8, 31-47.

Stahl, J., B. Bergstrom, and R. Gershon. 2000. “CAT administration of language placement examinations.” *Journal of Applied Measurement* 1, 292-302.

Syang, A., and N. B. Dale. 1993. “Computerized adaptive testing in computer science, assessing student programming abilities.” *SIGCSE Bulletin* 25, 53-57.

Thissen, D., and R. J. Mislevy. 2000. “*Testing Algorithms.” In H. Wainer, Computerized Adaptive Testing, A Primer 2*, 101-133. Mahwah, NJ, Lawrence Erlbaum Associates.

Triantis, D., and E. Ventouras. 2012. “Enhancing electronic examinations through advanced multiple-choice questionnaires*.”* In R. Babo, and A. Azevedo, *Higher Education Institutions and Learning Management Systems*, *Adoption and Standardization*, 178-198. Amsterdam, The Netherlands, Information Science Reference - IGI Global.

Tsiakas, P., C. Stergiopoulos, D. Nafpaktitis, D. Triantis, and I. Stavrakas. 2007. *“Computer as a tool in teaching, examining and assessing electronic engineering students.” Proceedings of the International Conference on ‘Computer as a Tool,’* 2490-2497.

Van der Linden, W. J., and C. A. W. Glas. 2000. *Computerized Adaptive Testing, Theory and Practice.* Heidelberg, Springer.

Ventouras, E., D. Triantis, P. Tsiakas, and C. Stergiopoulos. 2011. “Comparison of oral examination and electronic examination using paired multiple-choice questions.” *Computers and Education* 56, 616–624.

Wainer, H., and D. Thissen. 1993. “Combining multiple-choice and constructed-response test scores, Toward a Marxist theory of test construction.” *Applied Measurement in Education* 6, 103-118.

Wass, V., R. Wakeford, R. Neighbour, and C. Van der Vleuten. 2003. “Achieving acceptable reliability in oral examinations, an analysis of the Royal College of General Practitioners membership examination’s oral component.” *Medical Education* 37, 126–131.

Wei, H., and J. Lin. 2015. “Using Out-of-Level Items in Computerized Adaptive Testing.” *International Journal of Testing* 15, 50-70.