iMedPub Journals
http://journals.imedpub.com

Health Science Journal
ISSN 1791-809X

2015
Vol. 9 No. 2:8

Software-Assisted Identification and Improvement of Suboptimal Multiple Choice Questions for Medical Student Examination

Gerovasili Vasiliki[1], Filippidis Filippos T[2], Routsi Christina[1] and Nanas Serafim[1]

1 First Critical Care Medicine Department, Evangelismos Hospital, National and Kapodistrian University of Athens, Greece

2 School of Public Health, Imperial College London, United Kingdom

Correspondence: Gerovasili Vasiliki

a.icusn@gmail.com

First Critical Care Medicine Department, Evangelismos Hospital, National and Kapodistrian University of Athens**,** 45-47 Ipsilandou str., GR 106 75 Athens, Greece

Tel: 0030-210-7201918, 0030-6973036448

Fax: 0030-210-7244941

Abstract

Background: Multiple choice questions (MCQs) are often used to assess student achievement. Questions' content is mainly chosen by the tutor according to his judgment and experience. We aimed to develop an evaluation program of MCQs for medical students.

Method and Material: Specifically designed software was developed utilizing a database of all MCQs that were used to examine medical students. We evaluated 220 multiple choice questions used in a population of 497 students. For each question the Difficulty and Discrimination indices were calculated. The Discrimination index represents a question's discrimination ability -whether it has a high rate of success within high performing students. We evaluated 220 multiple choice questions used in a population of 497 students. A logistic regression model was tested to assess the association between Difficulty and Discrimination indices. Nineteen questions with Discrimination index lower than 0.20 were modified and given to 140 students.

Results: Out of the 220 questions, 37(16.8%) were of recommended difficulty while 30 (13.6%) were of "high difficulty - not acceptable" and 54 (24.5%) of "high facility- not acceptable". Seventy three questions were of excellent discrimination (33.2%), while 53 (24.1%) were of bad discrimination. Too easy and too difficult question were less likely to be of good/excellent discrimination (Odds ratio=0.18). The mean Discrimination index of the 19 questions that were modified improved significantly from 0.06 to 0.26 ($p<0.001$).

Conclusions: Choosing MCQs according to tutor's judgment only is not sufficient to create an objective evaluation system of students' achievement. The use of specifically designed software can help identify and improve flawed questions.

Keywords: Multiple-choice questions; Medical student exam software; Difficulty; Medical students

Introduction

Student training should involve, apart from lectures and practice, a credible method for the assessment of competence on the subject examined [1-3]. Bloom's taxonomy is an attempt to group educational objectives and is now used as a framework for the organization of curriculums and examinations [4]. Teachers usually spend considerable time in preparing lectures, but not as much when it comes to preparing written or oral examinations. They seldom have the required expertise to create reliable examination methods, as the majority of them have not undertaken training in examination methods [5].

Multiple choice questions (MCQs) are a very popular format

for student assessment. Its widespread use may be attributed to certain advantages, such as its versatility [6]. A large number of examinees can be assessed and a wide range of content can be covered [7,8]. It is also an objective method and evaluation is easy and quick. Disadvantages of multiple choice questions include the encouragement of cheating and the fact that they might be obscure or misleading [8]. They have been found to test lower-level but not higher-level cognitive functions, like synthesis [7-9]. Creating multiple choice items can also be complicated and time consuming for teachers [8].

When writing MCQs, certain guidelines should be followed [1,8], in order to create examinations appropriate for the evaluation of students. However, few studies have been conducted to evaluate the quality of MCQs either included in databases [7,9,10] or created by experienced teachers [5,7,11,12]. All of them have identified a large number of item writing flaws (IWF) which might have a significant effect on student performance. Even items in a highly prestigious journal were identified as flawed [13].

The objective of the present study was the development and appraisal of software which could be used for the evaluation and the optimization of multiple choice items used in undergraduate examinations for medical students.

Methods

Multiple choice questions

All MCQs that were used in student examinations of the Intensive Care course of the Medical School were used for the purposes of the present study. The Intensive Care course is a compulsory course taught in the fifth year (out of a total of six years of training) of the Medical School [14]. These questions had been written by members of the faculty and, subsequently, reviewed by two other members of the faculty. Questions were grouped by subject and organised in databases.

Software

Specifically developed, customized software was utilized for the examinations of the course and the evaluation of MCQs. The development of the software was a joint effort of the teachers and the company's programmers (Quicktesting, ANOVA consulting). All existing MCQs were imported in the software's database. The software provided data on the sample of students in which the question has been used, the percentage of students who have answered correctly (success rate) and the most recent examination date.

Questions in the database can be classified according to the date of the most recent examination in which they were used, the success rate of by subject. Selection criteria may also include Difficulty index and/or Discrimination index, which are automatically calculated for each question in the database.

Additionally, the software was designed to create tests and rearrange questions and answers within each test. Thus, alternative versions of the same test can be obtained. These versions were subsequently grouped and corrected automatically. Answer sheets were scanned by a specially designed scanner and correction was performed instantly. Test results and statistical data, both total and question-specific were calculated automatically by the software.

Difficulty and discrimination indices

Difficulty and Discrimination indices, which were automatically calculated by the software, may be used to evaluate multiple choice questions. Such indices are commonly used for evaluation purposes [10-12,15,16].

Answer sheets were scanned and a score was assigned to each examinee by the software. Subsequently, examinees were sorted by their individual score. Twenty-five percent of the students with the highest scores were included in the high performance group, whereas 25% of the examinees with the lowest scores comprised the low performance group. Thus, groups of high and low performance were defined. Calculation of Difficulty and Discrimination indices were based on these groups of examinees.

The difficulty index reflects the success rate in each question and ranges from 0 to 100 with higher values reflecting easier questions [10]. The difficulty index was calculated by the equation Difficulty index=(x+y)*100/n, where x=number of correct answers in the high performance group, y=number of correct answers in the low performance group and n=total number of examinees in the two groups.

The discrimination index is a measure of the question's ability to discriminate between students of high and low performance and it ranges from -1 to +1. Values closer to +1 indicate high discrimination. The discrimination index was calculated by the equation Discrimination index=2*(x-y)/n, where x=number of correct answers in the high performance group, y=number of correct answers in the low performance group and n=total number of examinees in the two groups.

Study design

Statistical reports for all available course MCQs were retrieved from the software and classified in groups, based on the values of the Difficulty and Discrimination indices. Based on the Difficulty index, questions were classified in five groups: high facility-non acceptable (≥70), acceptable facility (60-69); recommended difficulty (50-59); acceptable difficulty (30-49); high difficulty-non acceptable (<30) (**Table 1).** Based on the Discrimination index, they were classified in four groups: excellent discrimination (≥0.35); good discrimination (0.25-0.34); average discrimination (0.15-0.24); bad discrimination (<0.15) **(Table 2).**

Forty-five MCQs were selected from the database by the teachers and a test was created using the software. Nineteen questions were found to have Discrimination index lower than 0.20. These questions were reviewed by the researchers and item writing flaws were identified in all of them. They were modified accordingly and the test was used in end of term examination of 140 medical students. Subsequently, the test was corrected; indices and statistical data for the modified questions were automatically calculated.

Table 1 Grouping of 220 evaluated multiple choice questions according to Difficulty index

Group	Difficulty index	% (N)
High facility-non acceptable	≥70	24.5 (54)
Acceptable facility	60-69	18.2 (40)
Recommended difficulty	50-59	16.8 (37)
Acceptable difficulty	30-49	26.8 (59)
High difficulty-non acceptable	<30	13.6 (30)

Difficulty index reflects the success rate in the question (values from 0 to 100). Difficulty index=(x+y)*100/n, where x=number of correct answers in high performance group, y=number of correct answers in low performance group and n=total number of examinees in the two groups

Table 2 Grouping of 220 evaluated multiple choice questions according to Discrimination index

Group	Discrimination index	% (N)
Excellent discrimination	0.35 - 1	33.2 (73)
Good discrimination	0.25 - 0.34	21.8 (48)
Average discrimination	0.15 - 0.24	20.9 (46)
Bad discrimination	-1 – 0.15	24.1 (53)

Discrimination index reflects the ability of the question to discriminate between students of high and low performance (values from -1 to +1) Discrimination index=2*(x-y)/n, where x=number of correct answers in high performance group, y=number of correct answers in low performance group and n=total number of examinees in the two groups

Statistical analysis

Questions were grouped in a binary variable according to their Discrimination Index. Good and excellent discrimination categories were merged in one group, while average and bad discrimination categories were merged in the other group. We also created three groups, according to the Difficulty Index value. Recommended difficulty was used as the reference category; questions of acceptable facility and acceptable difficulty comprised the second category; and questions of unacceptable facility or unacceptable difficulty comprised the third category. A logistic regression model was fitted in the sample of 220 questions to assess the association between having good/excellent discrimination and the level of difficulty. Results are presented as Odds Ratio (OR) with 95% Confidence Intervals (95% CI).

To compare the proportions of questions with "bad discrimination" among those who were selected and corrected by the researchers before and after the amendments, a chi-square test was used. Mean values for continuous variables are presented with their Standard Deviation (SD) and range.

Results

A sample of 220 multiple choice questions was analyzed. All questions had been used in medical school examinations in the past. Each question had been answered by an average 224 (SD=105, range: 33-497) students in their fifth year of studies.

Mean question difficulty was 55 (SD=21, range: 3-99). Thirty-seven out of 220 questions (16.8%) were classified as of "recommended difficulty", 30 questions (13.6%) were classified as of "high difficulty-non acceptable" and 54 questions (24.5%) were classified as of "high facility-non acceptable" **(Table 1).**

Seventy-three out of 220 questions (33.2%) were classified as of "excellent discrimination", whereas 53 questions (24.1%) were classified as of "bad discrimination" **(Table 2).**

The majority of questions of "recommended difficulty" were found to be of "excellent" (18/37 or 48.6%) and "good discrimination" (8/37 or 21.6%). On the contrary, 15 out of 54 questions of "high facility-non acceptable" (27.8%) were classified as of "bad discrimination" and 17 more (31.5%) as of "average discrimination". Almost all questions of "high difficulty-non acceptable" were found to be of "bad" (56.7%) and "average discrimination" (33.3%). A visual examination of the data indicates that no questions of very low or very high difficulty showed optimal discrimination index **(Figure 1).**

Results from the logistic regression model showed that very easy or very difficult MCQs were much less likely to have good/excellent discrimination than MCQs of recommended difficulty (OR=0.18, 95% CI: 0.08-0.42) **(Table 3).** On the contrary, questions that were classified in the groups "acceptable facility" and "acceptable difficulty" were equally likely to be of good/excellent discrimination. When a similar model was fitted for excellent discrimination only, results were similar.

The 19 questions that were selected for modification had a mean Discrimination index of 0.06 (SD=0.09, range:-0.15 to 0.17). When assessed after modification, they were found to have mean Discrimination index of 0.26 (SD=0.11, range: 0.07 to 0.48). Fourteen out of 19 questions were initially of "bad discrimination", whereas only 2 of them remained in the same category after modification (p<0.001). Six out of the 19 questions were classified as of "good discrimination" and 4 as of "excellent discrimination" after modification.

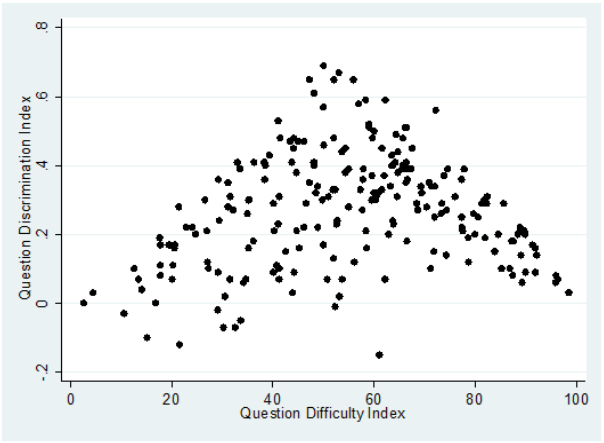

Figure 1 Scatterplot of Difficulty and Discrimination Indices for 220 Multiple Choice Questions

Table 3 Association between having good/excellent discrimination and level of difficulty

OR (95% CI)	
Recommended difficulty (ref)	1.00
Acceptable facility/acceptable difficulty	1.02 (0.45-2.34)
Unacceptable facility/unacceptable difficulty	**0.18 (0.08-0.42)**
OR=Odds Ratio; 95% CI=95% Confidence Interval	

Discussion

A major finding of the present study is that the employed software for the evaluation of multiple choice questions has revealed flaws in numerous items of the database, despite the fact that the evaluated items had been initially considered appropriate by experienced teachers. Many questions were found to be unacceptably difficult or easy and/or of bad discrimination. Additionally, an association between very low or very high difficulty and suboptimal discrimination of high versus low performing students was identified. Modifications made to questions that were identified thanks to indices calculated by the software, improved the discrimination significantly.

Evaluating multiple choice questions with the developed software is highly advantageous. Evaluation of the items is automated and rapid. Thus, identifying flaws and evaluating the quality of MCQs is considerably easier. Moreover, evaluation is possible before the announcement of the results; in case flaws are identified, results can be adjusted by eliminating flawed questions and not taking them into account based on examiners' judgment. Flawed items can be subsequently corrected in order to be used in future exams.

The existence of items of “high difficulty”, “high facility” and/or “bad discrimination” can be attributed to a number of reasons. Some of questions are clearly flawed, as they contain two or no correct answers. But there are several other common flaws that might not be so apparent, such as giving grammatical or logical cues; using absolute or very vague terms; repeating words; using an unusually long phrase for the correct answer; using different formats for numeric data; negative phrases; including “none of the above” or “all of the above” as options [17,18]. Flawed multiple choice questions are common [13,19]. Nevertheless, only a few studies have been performed to evaluate multiple choice questions that are given to medical students [5,9]. One study has evaluated the quality of multiple choice questions as poor [5] and another has found that 46% of the multiple choice items contained at least one item writing flaw [9]. In a study concerning accounting students, 75% of the questions that were reviewed had at least one guideline violation [20].

Flawed multiple choice questions might affect students’ performance. Removing flawed items from a test resulted in 10-15% of examinees who had failed achieving a pass grade [12]. Flawed questions have also been found to affect negatively high-achieving students [10]. Our study also showed that an inappropriate level of difficulty (too low or too high) was associated with lower ability to discriminate between high and low achieving students. Availability of software capable of calculating Difficulty and Discrimination indices facilitates the identification of such items, their modification and, therefore, the improvement of students’ evaluation.

Identification of flawed questions according to the Discrimination index was shown to be efficient, given that all of the 19 questions that were identified contained item writing flaws. Their modification resulted in a statistically significant decrease of the proportion of questions with “bad discrimination” ($p<0.001$). Ten of the 19 items were of “good” or “excellent discrimination” after modification. It seems that this software can be an effective tool to improve the quality of questions. Despite that, two of the modified questions remained in the group of “bad discrimination”. This is an indication that MCQs should be reviewed and reevaluated constantly. Computer software could be helpful in this aspect as well, as it provides constantly updated statistical information for each individual item.

Identification of flawed items and appropriate amendments might greatly reduce the proportion of unsuitable MCQs, but this may not be enough to construct an assessment of high quality. Questions should be linked to learning objectives and test students at a higher cognitive level [21,22]. Questions at the knowledge/recall level, even without flaws, might fail to discriminate between high and low achieving students. This is a possible explanation why a small number of the amended questions remained in the “bad discrimination” group.

One of the main disadvantages of multiple choice questions is that cheating might be encouraged. The developed software offers the option to rearrange questions and answers within the same test and create alternative versions of it. Through this process, cheating is discouraged and the quality of examinations is improved. The alternative versions are subsequently grouped and corrected simultaneously, without additional burden to the markers.

The development of the software was a joint effort by the teachers and the company's programmers and it required a considerable time investment by both teams. Time required for marking exams was improved substantially with the use of the software and it is expected that it will be reduced further in future reiterations, as problematic questions will be gradually removed from the pool. The software is commercially available (Quicktesting, ANOVA consulting) and its use for MCQs examinations may significantly reduce the time needed for exam marking, especially when a large number of students is being tested. However, we have not conducted a cost and benefit analysis; this should be addressed in future studies so as to determine affordability and potential savings for universities.

Teachers in medical schools have moral and professional responsibility to fairly and reliably assess their students’ performance. Examinations are frequently the only means to evaluate a student’s competence. Therefore, it is of major importance to have credible examination items. MCQs can be difficult to construct and time consuming, but it has been shown that training improves teachers’ ability to construct flawless multiple choice items. Nevertheless, training is not always available. In the light of this, software can be helpful for teachers to decrease the time needed for the construction of items, and, as shown in the present study, to identify flawed questions and modify them, thus improving the quality of tests. It may also serve as a tool for quality control and facilitate comparisons between different courses or disciplines [21].

The development of computer software for the optimization of examination with the use of multiple choice questions offered the opportunity to evaluate questions according to their difficulty and their discrimination. Thus, items of “excellent discrimination” and “recommended difficulty”, but also items of “bad discrimination” and “high difficulty” or “high facility” were identified. Through

this process, flawed items were found and modified quite successfully. Further evaluation of the questions in the database will enable the teachers to improve the quality of items and the assessment of students' performance.

Conclusions

MCQs are a very popular format for student assessment. However, choosing MCQs according to tutor's judgment only is not sufficient to create an objective evaluation system of students' achievement. The use of specifically designed software can help identify flawed questions as well as questions that are unacceptably difficult or easy, so that these flawed or inappropriate questions can be improved. It could therefore help teachers improve the quality of MCQs in order to better assess students' performance.

Declaration of Interest

The authors report no declaration of interest

References

1 Howley LD (2004) Performance assessment in medical education: where we've been and where we're going. Eval Health Prof 27: 285-303.

2 Anderson L, Krathwohl D, Airasian P, Cruikshank K, Mayer R, et al. (2001) A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom's taxonomy of educational objectives. Addison Wesley Longman New York.

3 Marzano RJ (2006) Classroom assessment & grading that work: Association for Supervision and Curriculum Development. Alexandria, VA.

4 Crowe A, Dirks C, Wenderoth MP (2008) Biology in bloom: implementing Bloom's Taxonomy to enhance student learning in biology. CBE Life Sci Educ 7: 368-381.

5 Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, et al. (2002) The quality of in-house medical school examinations. Acad Med 77: 156-161.

6 Schuwirth LW, van der Vleuten CP (2004) Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 38: 974-979.

7 Tarrant M, Knierim A, Hayes SK, Ware J (2006) The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Educ Today 26: 662-671.

8 Farley JK (1989) The multiple-choice test: writing the questions. Nurse Educ 14: 10-12, 39.

9 Masters JC, Hulsmeyer BS, Pike ME, Leichty K, Miller MT, et al. (2001) Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. J Nurs Educ 40: 25-32.

10 Tarrant M, Ware J (2008) Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 42: 198-206.

11 Downing SM (2002) Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? Acad Med 77: S103-S104.

12 Downing SM (2005) The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ Theory Pract 10: 133-43.

13 Stagnaro-Green AS, Downing SM (2006) Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. Med Teach 28: 566-568.

14 Nanas S, Gerovasili V, Poulaki S, Bouhla A, Tripodaki E, et al. (2008) Optimization of multiple choice examinations with the use of specifically designed software. Archives of Hellenic Medicine 25: 781-785.

15 DeSantis M, McKean TA (2003) Efficient validation of teaching and learning using multiple-choice exams. Adv Physiol Educ 27: 3-14.

16 Tripp A, Tollefson N (1985) Are complex multiple-choice options more difficult and discriminating than conventional multiple-choice options? J Nurs Educ 24: 92-98.

17 Al-Faris EA, Alorainy IA, Abdel-Hameed AA, Al-Rukban MO (2010) A practical discussion to avoid common pitfalls when constructing multiple choice questions items. J Family Community Med 17: 96-102.

18 Brunnquell A, Degirmenci U, Kreil S, Kornhuber J, Weih M (2011) Web-based application to eliminate five contraindicated multiple-choice question practices. Eval Health Prof 34: 226-238.

19 Nedeau-Cayo R, Laughlin D, Rus L, Hall J (2013) Assessment of item-writing flaws in multiple-choice questions. J Nurses Prof Dev 29: 52-57.

20 Hansen J, Dexter L (1997) Quality Multiple-Choice Test Questions: Item-Writing Guidelines and an Analysis of Auditing Testbanks. J Econ Bus 73: 94-97.

21 Ware J, Vik T (2009) Quality assurance of item writing: during the introduction of multiple choice questions in medicine for high stakes examinations. Med Teach 31: 238-243.

22 Vanderbilt AA, Feldman M, Wood IK (2013) Assessment in undergraduate medical education: a review of course exams. Med Educ Online 18: 1-5.