HOSPITAL CHRONICLES 2011, 6(2): 79–83

ORIGINAL ARTICLE

# The Examination of Accreditation for Foreign Medical Graduates in Greece: Evaluation of the Multiple Choice Question Format Using Difficulty and Discrimination Indices

Vassiliki Kostopoulou, MD, Csilla Zafiriou, MD, John Lymveos, MD, Chrysanthi Trikka, MD, Savvas Toumanidis, MD

*Department of Clinical Therapeutics, Medical School, University of Athens, "Alexandra" Hospital, Athens, Greece*

**KEY WORDS:** *accreditation; difficulty index; discrimination index; multiple choice questions; foreign medical graduates*

**LIST OF ABBREVIATIONS**
DfI = difficulty index
DsI = discrimination index
MCQ = multiple choice question

*Correspondence to:*
Savvas Toumanidis, MD
"Alexandra" Hospital
Department of Clinical Therapeutics
80 Vassilisis Sofias Avenue & Lourou Street
Athens 115 28, Greece
e-mail: stouman@otenet.gr

*Manuscript received December 24, 2010; Revised manuscript received January 15, 2011; Accepted January 21, 2011*

**ABSTRACT**

**BACKGROUND:** The multiple choice question (MCQ) format is the most commonly used written assessment technique for the accreditation of foreign medical graduates in Greece.

**OBJECTIVES:** To evaluate the appropriateness of the range of the multiple choice questions used in the examination for the accreditation of foreign medical graduates in Greece, and to compare the performance among foreign medical graduates, Greek medical students and interns (medical doctors during their residency).

**METHODS:** Twenty-six items from the internal medicine question paper and 24 items from the surgery question paper were randomly selected from an MCQ format used in one assessment. For these items discrimination and difficulty indices were calculated for separate groups of candidates and volunteer participants (53 medical students and 30 interns). Comparisons were made between group scores, first considering the whole questionnaire as a single entity and then using scores for each discipline calculated separately.

**RESULTS:** A significant number of "inappropriate" questions were included in the examination. Surgery questions were more candidate-oriented, given the best range of acceptable difficulty index values for that group of participants, while internal medicine questions proved to be more appropriate for medical students. Furthermore, comparisons of groups performed using a total score over the whole range of the two disciplines revealed a significantly better performance of interns compared with students ($p<0.001$), whereas comparisons performed separately for each discipline revealed no significant difference between interns and students in surgery scores but a significant difference in internal medicine scores ($p<0.001$).

**CONCLUSION:** These findings suggest the importance of the evaluation of the MCQs before using them in examinations, aiming at revising inappropriate questions. In order to evaluate the participants' performance, calculation of scores across separate disciplines is proposed, since it is less likely to be biased towards good performance in the questions of one discipline.

Conflict of interest: None declared

## INTRODUCTION

Certifying evaluation is designed to protect society by preventing incompetent personnel from practicing medicine. The multiple choice question (MCQ) format is the most commonly used written assessment technique for such tests, because of its many positive psychometric characteristics, its long history of research evidence, its versatility in testing most cognitive knowledge, its relative (apparent) ease to write, store, administer and score, and its continued use by the highest state examinations in medical education.[1,2] Multiple choice questions are able to test a number of skills, such as understanding, reasoning, data interpretation and problem solving, in addition to the recall of factual knowledge. They are reliable, discriminatory, reproducible and cost-effective. In comparison with other examination methods – such as oral, practical, essay examinations – a very important characteristic of the MCQ format is the possibility of its evaluation.[2,3] The classical pedagogical analysis of the MCQ uses the difficulty and discrimination indices,[4] but there are some newer ways to assess test question value and validity using computer generated matrices.[5] The *difficulty index* measures the easiness (or difficulty) of a test question, whereas the *discrimination index* indicates how effectively a question discriminates between "high/good" and "low/bad" examinees.[6] Discrimination indices express whether there is any lack of correlation between the performance of candidates in answering an individual test question and their overall performance in the whole paper.[7] According to these tests a critical evaluation of each question is performed, enabling a given question to be retained, revised or rejected.

The examination for the accreditation of foreign medical graduates who desire to practice medicine in Greece is an MCQ assessment consisting of internal medicine and surgery questions. The peculiarity of this examination is that a pre-test evaluation of the MCQ is ruled out by the requirement of strict secrecy and confidentiality. The purpose of the present study was 1) to evaluate the appropriateness of the MCQ tests for the examination of accreditation for foreign medical graduates in Greece, using the difficulty and discrimination indices, and 2) to compare the success rate between foreign medical graduates and Greek medical students as well as interns in Internal Medicine.

## METHODS

The examination of accreditation for foreign medical graduates in Greece is performed twice annually. For the examination two separate question papers are prepared on two occasions, containing 100 internal medicine and 100 surgery questions, each consisting of a stem and five completing phrases. The stem is read in turn with each of the five phrases, and boxes are provided for the candidate to mark each of the five resulting sentences as "true", "false", or "don't know". True items marked true contribute to a correct score while false items marked true contribute to an error score. Items marked "don't know" do not contribute to an error score. The error score is deducted from the correct score to give the candidate's final score for each discipline. Scoring is performed automatically by a specialized computer-assisted program for MCQ evaluation. There is a requirement that candidates should pass each of the two disciplines.

The evaluation of the MCQ in this study was performed after a random selection of 26 items (26 questions each with five true/false responses) from the internal medicine question paper and 24 items (24 questions each with five true/false responses) from the surgery question paper. Overall, 118 surgery and 130 internal medicine questionnaires were chosen from among those completed by the candidates who participated in one assessment. The selection was made in order to maintain the percentages of success/failure observed after the announcement of the results, but since candidates' names were not visible for reasons of confidentiality, the internal medicine and surgery question papers answered by a given candidate could not be correlated or compared. For each discipline, graduates' ranking in relation to the score calculated from the randomly selected items was performed, proceeding from the highest to the lowest score. The participants then were categorized into three groups based on their scores. These included those with scores lying in the lowest quartile, the highest quartile, and the remainder, and constituted a choice distribution table. For each item, this table was then utilized to calculate the indices of difficulty and discrimination. In order to quantify the difficulty of the questions for each discipline, the *difficulty index (DfI)* was calculated using the formula: DfI (%) = [(H+L)/N]*100; where H: correctly selected true items in the high score group, L: correctly selected true items in the low score group and N: total number of candidates in the two groups.[6] The reason for measuring item difficulty is to choose items of a suitable difficulty level which will help in assessing as accurately as possible each candidate's level of knowledge. A difficulty index of 30% - 70% is considered acceptable because it is very likely to be reliable as regards its internal consistency or homogeneity. Questions with difficulty index greater than 70% are considered too easy while ones with index less than 30% are considered too difficult. For each item, the choice distribution table was then used to calculate the discrimination index (DsI) given by the formula: [6]

$$DsI = 2*(H-L)/N$$

A discrimination index of 0.25 or greater is considered good, one of 0.16 - 0.24 indicates that the question needs to be reconsidered, while one of less than 0.15 is poor and the question should be discarded.

Subsequently, both question papers were voluntarily an-

swered by 53 undergraduate students of the Athens Medical School and 30 interns in an internal medicine department. Discrimination and difficulty indices were also calculated for each of these groups using the methodology mentioned above.

### STATISTICAL ANALYSIS

The equality of proportions of unacceptable questions between different groups was checked using two-sample tests for proportions. Comparisons of mean scores among interns', students' and candidates' question papers were performed using a two-sided t-test or analysis of variance (ANOVA). Bonferroni post-hoc analysis was used to test for multiple comparisons. Data are presented as mean ± standard error (SE). The level of statistical significance was set at 0.05 in all cases.

## RESULTS

### DIFFICULTY AND DISCRIMINATION INDICES

After the difficulty index for internal medicine and surgery questions was calculated for each group of participants, 46%, 42% and 27% of the internal medicine questions, and 50%, 71% and 29% of the surgery questions were found to be inappropriate (too difficult or too easy) for medical students, interns and foreign candidates, respectively. Specifically, surgery questions were significantly more inappropriate for interns than for candidates ($p<0.04$). Internal medicine questions were easier for interns than for students or foreign candidates at a statistically significant level ($p<0.03$ and $p<0.001$, respectively) (Table 1).

Calculation of the discrimination index revealed that 19%, 58% and 42% of the internal medicine questions, and 46%, 50% and 0% of the surgery questions were not discriminative for medical students, interns and foreign candidates, respectively. Internal medicine questions were significantly more discriminative for medical students than for interns or foreign candidates, while surgery questions were significantly more discriminative for foreign candidates than for students or interns (Table 2).

### COMPARISON BETWEEN FOREIGN CANDIDATES-STUDENTS-MEDICAL DOCTORS

A comparative study of the mean scores among foreign candidates, students and interns revealed some interesting

**TABLE 1.** Comparisons of the quality of internal medicine and surgery questions for students, interns and foreign candidates, based on the Difficulty Index (DfI).

| | Difficulty Index (DfI) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Students (I) | | Interns (II) | | Candidates (III) | | I/II | | I/III | | II/III | |
| Internal Medicine | N | (%) | N | (%) | N | (%) | z | p | z | p | z | p |
| **Not acceptable*** | 12/26 | (46) | 15/26 | (42) | 7/26 | (27) | 0.29 | 0.77 | 1.42 | 0.15 | 1.13 | 0.25 |
| **Too Easy** | 7/12 | (58) | 14/15 | (93) | 2/7 | (29) | -2.16 | 0.03 | 1.22 | 0.22 | 3.13 | 0.001 |
| **Surgery** | | | | | | | | | | | | |
| **Not acceptable** | 12/24 | (50) | 17/24 | (71) | 10/24 | (42) | -1.48 | 0.13 | 0.55 | 0.57 | 2.02 | 0.04 |
| **Too Easy** | 7/12 | (58) | 12/17 | (71) | 4/10 | (40) | -0.72 | 0.46 | 0.84 | 0.40 | -1.58 | 0.11 |

*not acceptable (too difficult OR too easy) if DfI <30% or DfI >70%

**TABLE 2.** Comparisons of the quality of internal medicine and surgery questions for students, interns and foreign candidates based on the Discrimination Index (DsI)

| | Discrimination Index (DsI) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Students (I) | | Interns (II) | | Candidates (III) | | I/II | | I/III | | II/III | |
| Internal Medicine | N | (%) | N | (%) | N | (%) | z | p | z | p | z | p |
| **Not discriminative**** | 5/26 | (19) | 15/26 | (58) | 11/26 | (42) | -2.89 | 0.004 | -1.8 | 0.07 | 1.15 | 0.24 |
| **Surgery** | | | | | | | | | | | | |
| **Not discriminative** | 11/24 | (46) | 12/24 | (50) | 0/24 | (0) | -0.27 | 0.78 | 3.78 | <0.001 | 4.00 | <0.001 |

**not discriminative if <0.15

findings. When the comparison was performed using a total score (available only for the students' and the interns' group) over the whole range of the two disciplines, the two groups differed significantly from each other, interns presenting significantly higher scores given their expected better performance in internal medicine questions (t-test = -4.06, $p<0.001$). When comparison was performed separately for each discipline, mean scores differed significantly among all groups for both internal medicine (ANOVA: $F = 43.14$, $p<0.001$) and surgery (ANOVA: $F = 16.59$, $p<0.001$) question papers. In this last case, though, the Bonferroni post-hoc analysis demonstrated that there was no statistically significant difference between interns and medical students, whereas foreign candidates were found to perform significantly worse in both disciplines. The above results are shown in Figures 1, 2 and 3.

## DISCUSSION

In this study, application of the difficulty and discrimination indices revealed a significant number of inappropriate (too difficult, too easy, or not discriminative) questions included in the examination for accreditation of foreign medical graduates. The difficulty index did not exhibit any statistically significant differences in proportions of unacceptable internal

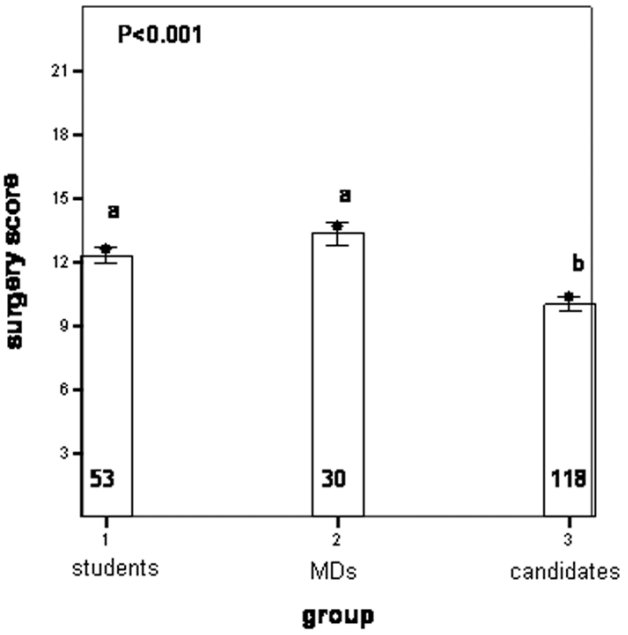


**FIGURE 2.** Mean ***surgery*** scores ± SE from ANOVA-adjusted pooled data of the three groups. Different letters above bars represent groups that were significantly different according to the ANOVA Bonferroni post-hoc analysis ($p<0.001$). Numbers in bars represent sample size.

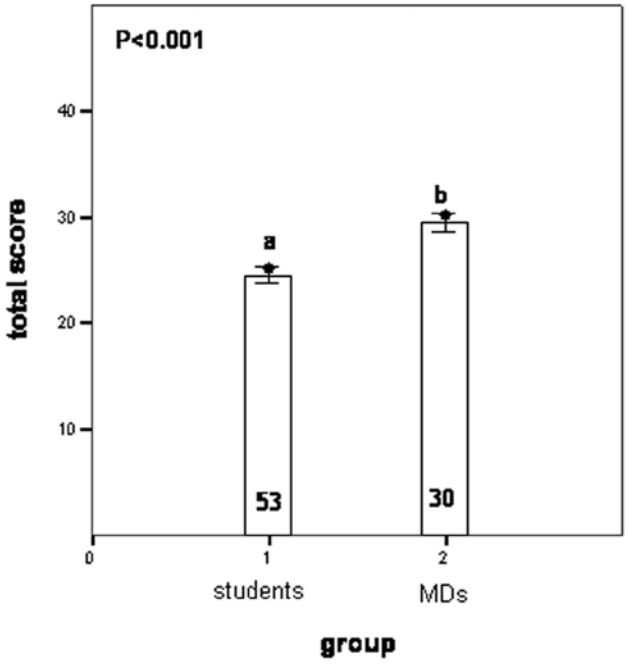


**FIGURE 1.** Mean ***total*** scores ± SE as derived from the t-test comparing the two groups. Different letters above bars represent significantly different groups ($p<0.001$). Numbers in bars represent sample size.

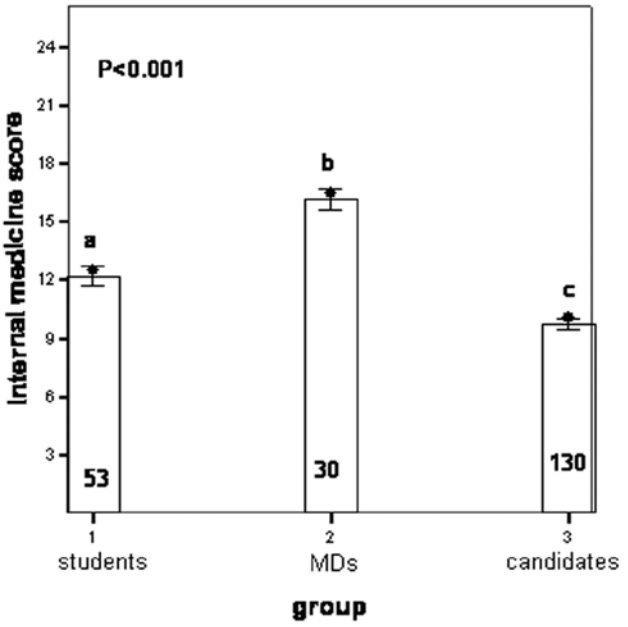


**FIGURE 3.** Mean ***internal medicine*** scores ± SE from ANOVA-adjusted pooled data of the three groups. Different letters above bars represent groups that were significantly different according to the ANOVA Bonferroni post-hoc analysis ($p<0.001$). Numbers in bars represent sample size.

medicine questions among groups (although the proportion of questions rejected due to easiness was significantly higher for interns, as expected). Surgery questions in this case proved to be more foreign candidate-oriented, given the best range of acceptable difficulty index values for that group of participants. One may hypothesize that this may be possibly due to the bad performance of medical students some of whom had probably not yet taken the surgery course exam at the time of the investigation, although there is no way to verify as to how many of the medical students had not yet done so. On the other hand, the discrimination index indicated that internal medicine questions were more discriminative for students than for candidates, while surgery questions were more discriminative for foreign candidates.

Comparisons of mean scores performed using a total score for both disciplines and then separately for each discipline gave different performance results. These findings indicate that papers analyzed across separate disciplines are less likely to be biased towards good performance of the questions in one discipline. Such a bias would result in successful candidates being those who were good in one discipline, even though they might not be good in the other.

### LIMITATIONS

This study has several limitations. A limited number of interns and medical students were interested in completing the question papers voluntarily (perhaps due to lack of incentives), therefore group sizes are disproportionate. On the other hand, foreign candidates were more highly motivated in completing the question paper than the other groups, therefore less likely to make careless mistakes. Moreover, calculation of a total score for both disciplines in the foreign candidates group was not feasible, given that internal medicine and surgery question papers for the same person could not be collated. The fact that surgery questions proved to be more candidate-oriented cannot be adequately explained, as there is no way to verify as to how many of the medical students had not yet taken the surgery course exam at the time of the investigation. Some may also wonder about the need for using a "don't know" item in an exam, since correct or incorrect responses to a question in some instances do not necessarily imply that the examinee knows. Finally, the conclusions deduced from the evaluation of a single session's MCQ format should not be generalized. Results from analysis of MCQ formats from various years' sessions would reflect more accurately the appropriateness of the MCQs used in the examination of accreditation for foreign graduates in Greece.

### PERSPECTIVE

Passing or failing achievement tests in medical education has serious consequences for examinees and, ultimately, for patients.[1] Therefore, the examiners have the responsibility to evaluate foreign candidates' qualifications correctly, using objective examination tools that are capable of reflecting the truth. They are also compelled to collect and present strong evidence that the test which they are using measures what it is intended for, and that the inferences drawn from test scores are more or less accurate and defensible.[1] Examinations using objectively scored test formats are developed and used extensively by faculties for local medical education use. It is essential though, that each institution should aim at controlling or eliminating all potential sources of bias during the construction of question papers as well as the interpretation of scores. An effort should be made to obtain the cooperation of all involved, and a study based on multiple evaluations is needed so that the conclusions reached are really representative.

## CONCLUSION

In conclusion, given the importance of the examination of accreditation for foreign medical graduates (only candidates who pass both disciplines are allowed to practice medicine in Greece), evaluation of the effectiveness of MCQs before using them in examinations is necessary in order to revise inappropriate questions. However, significant practical problems should be resolved in this case, such as the maintenance of secrecy and confidentiality. Furthermore, this paper shows that in order to avoid compensation for poorer performance in one topic by better performance in another in the case of examinations consisting of various component disciplines, evaluation of participants' scores should be done for each discipline separately, with a prerequisite of a minimum score for each component discipline before a candidate is allowed to pass.

## REFERENCES

1. Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct under representation. *Adv Health Sci Educ Theory Pract* 2002;7:235-241.
2. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ* 2004;38:974-979.
3. Auewarakul C, Downing SM, Jaturatamrong U, Praditsuwan R. Sources of validity evidence for an internal medicine student evaluation system: an evaluative study of assessment methods. *Med Educ* 2005;39:276-283.
4. Sim SM, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Ann Acad Med Singapore* 2006;35:67-71.
5. DeSantis M, McKean TA. Efficient validation of teaching and learning using multiple choice exams. *Advan Physiol Edu* 2003;27:3-14.
6. Guilbert JJ. Educational Handbook for Health Personnel, 6th edition. Geneva: World Health Organization; 1987
7. Hobsley M. Counting apples with oranges: a limitation of the discrimination index. *Med Educ* 1999;33:192-196.