A new methodology for comparison of three-test exam techniques in medical students

Nayer Rassaian, M.D., PhD

[1]Professor of Physiology, Shaheed Beheshti University of Medical Sciences and Health Services

ABSTRACT

Background: *In studies on validities of different methods for student assessment, each technique has been independently evaluated, but literature confirms the use of combination of different methods of assessment.*

Objectives: *Searching for negative and positive aspects of assessment tools, and introducing the more preferable technique.*

Methods: *In this descriptive-analytical research, 1372 medical students were tested using three types of questions; short-answer, multiple-choice, and true-false, so that the result of each method can be compared with those of the other two for each student. Nine classes from different curricula were randomly selected and in each of them, 30 questions for 10 selected topics (total 270), were distributed.*

Results: *The students' score analysis after excluding questions with a discrimination index of less than 0.3, showed that the most valid assessment tool was the short-answer questions (r0.685). The kappa coefficients of the short-answer method and the other two were in the "fair" range of agreement. The highest coefficient of contingency is between the multiple-choice and true-false questions (0.402). The percentage of correct answers is the lowest in short-answer questions (53%), and the highest in true-false questions (65%). The higher percentage of incorrect answers in multiple-choice (32%) compared with the true-false (26%) questions can be the result of the students being confronted with alternative choices in multiple-choice question, which apparently encourages them to choose one without using the relevant knowledge.*

Conclusion: *It is recommended to use short-answer and true-false questions as the main components of examination, instead of multiple-choice question alone, so that student's learning and recall would be tested and students would not be misled.*

Keywords: *MEDICAL STUDENT ASSESSMENT, SHORT-ANSWER, TRUE-FALSE, MULTIPLE-CHOICE, EXAMINATION QUESTIONS.*

Journal of Medical Education Spring *2004; 5(1);3-10*

Introduction

Having information about examination technique is an important part of an individual's preparation for the exam. In 1998, Hammond et al. performed a study on the candidates for the fellowship of the Royal College of Anesthetists in the United Kingdom, where multiple choice questions were part of the assessment process in both preliminary and final examinations. A group of candidates attended a pre-examination revision course. The study concluded that failure to appreciate the relationship between knowledge and assessment technique may significantly affect a candidate's performance in the examination (1); therefore, it can be said that a major part of students' preparation depends on their previous conception of the examination method. However, in the case of badly designed or repeated questions in successive exams, the knowledge of usual exam techniques leads students to study the question patterns rather than to think about the scientific problems at hand.

According to the study conducted by Mavis et al. although currently a variety of competency assessments are used, multiple-choice questions remain a core assessment method (2). In Irans' medical schools the usual way of examination - besides the oral exams during the clinical clerkships- is multiple choice questions. They are used most often because of the crowded classes,

especially during the first five years with basic sciences, physiopathology, and clinical courses. During the clinical clerkships, students are divided into groups in each ward, and then they are evaluated by oral exams and essays in addition to multiple-choice tests.

The current form of student knowledge assessment had some shortcomings.

In designing such questions, consideration should be given to the cognitive educational objectives which should be observable and measurable (3). Regarding different standard test exams, and their positive aspects which are discussed in the educational text books, however the necessary principles of designing such questions may not be adhered to(4). The questions received for American Board of Medical Examination which are mostly MCQS, are chosen after deep and critical review and long analytical discussions, by the National Board of Medical Examiners and the rest will be discarded (5). Difficulty in designing questions may lead to improper alternatives and therefore inappropriate student assessment. The reason why students claim that the MCQs in ordinary university exams are not a proper way of evaluating their knowledge properly, might be due to the above mentioned points.

In addition, the ability to answer questions by studying those of previous exams has caused an alarming reduction in class attendance, depriving the students of exchanging and analysing information, and exploratory thinking, hence quality learning (6,7). In fact, taking these points into consideration during many years of teaching by one of the authors (Rassaian, N.) was the initiative for present research to find a better way to evaluate test exams.

Due to subsequent effects of evaluation methods on studying techniques and usage of learned material, a change is felt necessary. The best way to benefit from multiple-choice method is its combination with other assessment techniques (8). Accordingly, the present study was designed to compare three different assessment methods in which each selected topic was assessed by three methods in each student. The variables were the number of correct, incorrect, and blank answers, the time required for marking the exam papers, and the students' scores. The final goal was to determine the preferred assessment technique obtained from this new methodology.

Materials & methods

This study was designed as a descriptive-analytical research and was conducted using the second semester exams of 1993 medical students from Shaheed Beheshti and Iran Universities of Medical Sciences and Health Services in the three curricula of basic sciences, physiopathology, and clinical sciences. These are sequential periods in a 7 year medical university program, consisting respectively of 4, 2, and 5 semesters, before the final 18 months of internship.

The selection was based on the university's agreement and the professor's interest to participate in designing exam questions according to the approved research program. After the universities' authorities gave the authorizations, three course examinations were selected randomly from each of the three curricula using the official examination schedule of the universities, i.e. 9 courses overall.

For each of ten selected topics in each course, questions were designed by using the three methods of multiple-choice, true-false, and fill in the blanks with appropriate word (short- answer), i.e. 30 questions overall for each course. The questions were of Taxonomy type 2, where the answers of each method of questioning were not hinted at in the body of questions of other methods. The question key was given by the question designer.

The questions for all nine courses were typed in two pages and the students were reminded that points would be deducted for incorrect answers for the first and second ten questions (multiple-choice and true-false, respectively) and that the third ten questions (short-answer) had no negative grades. The order of the questions was such that the first multiple-choice, the first true-false, and the first short-answer questions were about the same scientific topic and this order was repeated for all ten topics. The questions were included in the formal final exam, attached as last two pages. All of the collaborative professors wanted to consider

the score of this part of the exam as additional points or part of the students' final grade.

After each examination, thirty questions designed for this research were marked using the answer keys by one of our colleagues and were rechecked afterwards. They were asked to write the number of incorrect answers, the number of unanswered questions, and the marking duration for the ten questions of each type.

The students' scores were delivered in the shortest possible time to the related professors and, according to the given authorizations, the exam papers were considered as research documents and were kept by the research director for statistical analysis. Performing the research related exams did not interfere in any way with the universities' schedules or policy of examinations.

The question design procedure was as follows:

- Codes 1 to 10 for multiple-choice
- Codes 11 to 21 for true-false
- Codes 21 to 30 for short-answer

The content of the questions were designed so that codes 1, 11, and 21 were about one scientific topic, and 2, 12, and 22 from another one, and so on.

Results

General description of the population under study is shown in Table 1.

TABLE 1. Population under study, based on curriculum, name of course, number of students, and name of universities

	Class & Course	No. of Students	Medical University
Basic Sciences	1- Renal physiology & acid-base balance	256	Shaheed Beheshti
	2- Immunology	96	Iran
	3- Gastrointestinal physiology	181	Iran
Physiopathology	4- Pathology (organ systems)	73	Shaheed Beheshti
	5- Gastrointestinal	128	Shaheed Beheshti
	6- Rheumatology	121	Iran
Clinical Sciences	7- Gynecology & Obstetrics	136	Shaheed Beheshti
	8- Neurosurgery	232	Shaheed Beheshti
	9- Neurology	149	Shaheed Beheshti

The description of the questions in terms of difficulty index, and facility index is shown in Table 2. In each curriculum, the maximum number of difficult questions (<30%) were in basic sciences (25190), and the minimum number were in clinical period (6190), on the other hand the minimum number of easy questions (>70%) were in basic sciences (1 8190), and the maximum were in clinical phase (45190).

TABLE 2. Frequency distribution of questions with their difficulty index (DI) and facility index (FI) in three different curricula in medial Schools versus assessment techniques. (DI<30% & FI>70%)

	Basic Sciences		Physiopathology		Clinical Sciences		Total	
	<30%	>70%	>30%	>70%	>30%	>70%	<30%	>70%
Multiple-choice	5	8	7	8	2	13	14	29
True-false	9	7	3	7	0	22	12	36
Short-answer	11	3	7	6	4	10	22	19
Total	25	18	17	21	6	45	48	84

Since the quality of the questions is an interfering variable in analysing the responses, the discrimination index of each of the 270 questions (90 questions for each method) was calculated and a total of 39 questions with a discrimination index of less than 0.3 were considered unacceptable. Certainly, when it was necessary to discard a question designed for one method the related questions by the other two methods were also discarded, because in this research methodology, each method is the control for questions of that topic asked by the other methods and omitting one leads to the other two's omission. Therefore, another 51 questions were omitted. Thus, a total number of 90 questions were discarded, and all the statistical analyses were done on the remaining 180 questions, i.e. 60 for each method (Table 3).

TABLE 3. Frequency distribution of exam questions and all of the answers in each method after omitting questions with DI<0.3 for the 9 classes in the three curricula.

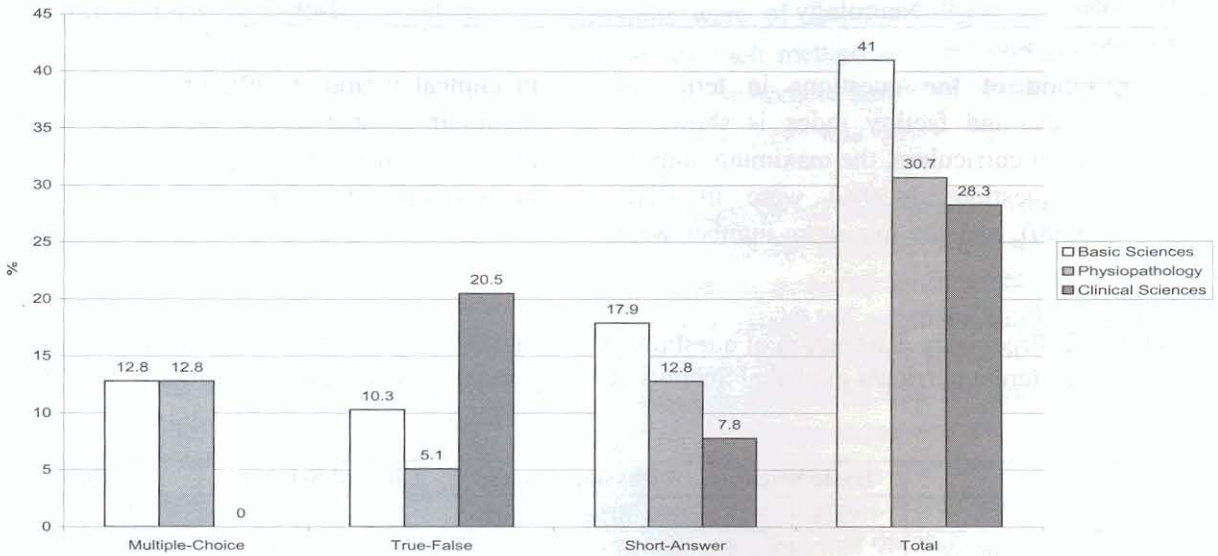
	Basic Sciences			**Physiopathology**			**Clinical Sciences**			**Total**
Class Number©	1	2	3	4	5	6	7	8	9	
No. of questions	7	5	6	7	7	7	7	8	6	60
No. of all answers	1792	480	1086	511	896	847	952	1856	894	9314

©Class numbers are indicative of the courses in table 1

The description of the questions with discrimination index less than 0.3 is shown in Figure 1 Clinical phase had the maximum number by true-false method (20.5%), with no unacceptable MCQS.

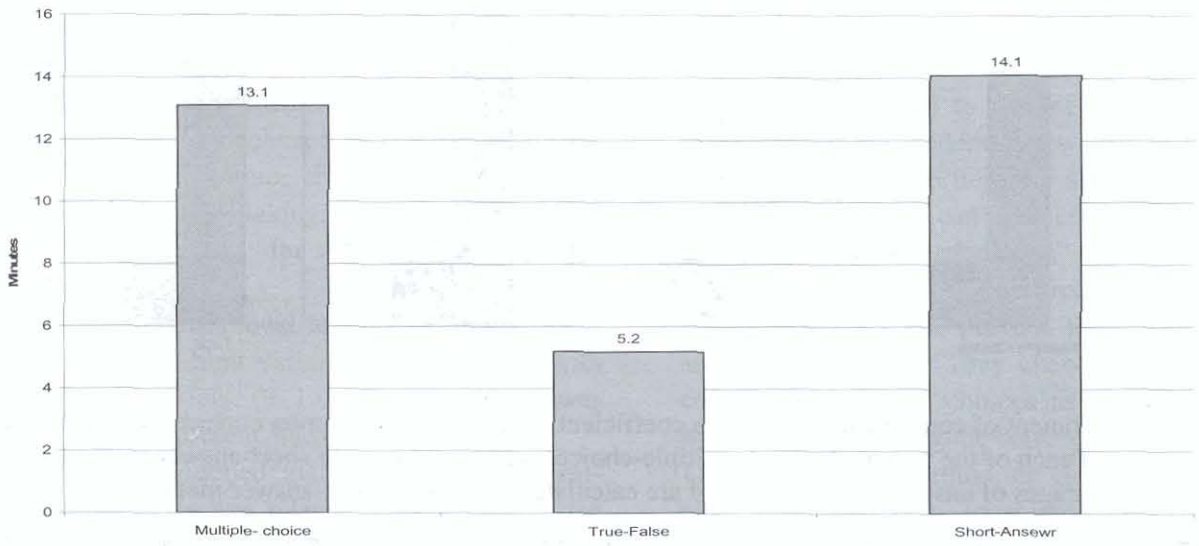
Average time for correcting papers was calculated for each method, which was compared with the Duncan's multiple range and t test.

FIGURE 1. Percentage of questions with DI<0.3 in three different curricula in medical schools by assessment techniques



As shown in Figure 2 the true-false method required the least amount of time (5.2 + 2.3 minutes) and statistically it had a significant difference with the multiple-choice method (13.1+ 5.4 minutes), which is significantly shorter than short-answer method (14.1 + 7.9 minutes), which took the longest duration,as expected (P<0. 000 1).

FIGURE 2. Mean of marking duration of each ten questions relevant to its assessment techniques



Studying the number of correct, incorrect, and blank answers in the remaining 60 questions for each method of assessment (Figure 3) showed that, although no negative grades for incorrect answers were given, the short-answer method had the greatest number of incorrect answers (39%) and that such questions were relatively difficult for the students. However, it is notable that incorrect answers were 26% in true-false questions and 32% in multiple-choice questions. True-false questions had also the greatest number of correct answers (65%), which can not be interpreted based on its 50% chance of guessing because the greatest number of unanswered questions were also seen in this type of questions (9%). The results of kappa test and the coefficient of contingency for comparing the degree of agreement between the short-answer and each of the other methods is shown in Table 4, which shows that the number of common correct answers between MCQs and short-answers is equal to that of true-false and short-answer (78%). These tests were also done for evaluating agreement between multiple-choice and true-false methods resulting number of common correct answers 4227, common incorrect answers 1212, and common blanks 335 with kappa coefficient 0.269 (p< 0. 00000 1), which showed the highest coefficient of contingency (0.402).

After discarding the questions with DI < 0.3, the students' scores were calculated on the basis of 20, and by factor analysis of the resulting scores, a related factor was obtained, the common element of its constituents is considered to be student's knowledge. It is most probable that the variable representing a higher correlation with this factor will have a higher importance in evaluating their knowledge. The above analysis showed that, in relation to other two methods, the short-answer had the highest correlation (r = 0.685), and the multiple choice had a very close correlation to it (r = 0.677).

FIGURE 3. The number of correct, incorrect, and blank answers in the remaining 60 questions

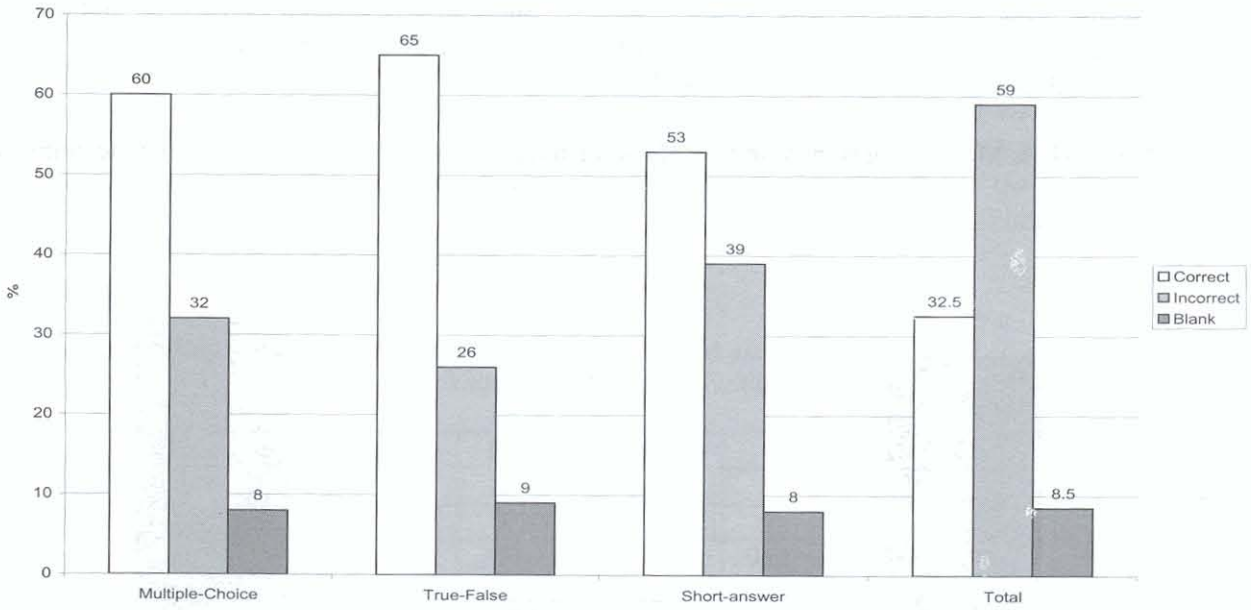


TABLE 4. Coefficient of contingency and kappa coefficient for comparing common correct, incorrect, and blank answers of each of the two methods of multiple-choice and true-false, with short-answer method. The percentages of answers in each method are calculated relative to short-answer method.

	Correct (%)	Incorrect (%)	Blank (%)	Contingency Coefficient	Kappa	P
Multiple-choice with Short-answer	3801(78)	1853(51)	214(28)	0.395	0.330	0.000001
True-false with short-answer	3815(78)	1426(39)	253(33)	0.343	0.253	0.000001

The correlation of true-false tool appeared in the third order (r=0. 5 60). Also, since the comparison of the assessment methods was our the objective, it seemed necessary to compare the mean of students' scores from acceptable questions obtained from each technique - after those with a Dl< 0.3 was discarded - with the related mean of raw scores, which provided a very close results (Table 5).

TABLE 5. Mean of raw scores of 1372 medical students with the three assessment methods before and after omitting questions with DI<0.3. The calculated correlation of each method with the student's grades, after omitting the questions, is also presented*

	Score Mean**		r
	Before	After	
Short-answer	10.2	10.5	0.685
Multiple-choice	11.7	11.9	0.677
True-false	13.0	12.9	0.560

Discussion

So far the validity of test exam questions designed by a method has been determined independent from other methods. In this study student's knowledge of each scientific fact was assessed by three combined methods: multiple-choice, true-false, and short-answer. Although the students are unaccustomed to short-answer questions in comparison with the other two, this creative way of assessment was considered as the most valid one. As a result, it was selected as the control group. And in order to reduce the amount of probable guessing and chance taking, it seemed necessary to use negative marks for MCQs and true-false questions.

According to educational text books, short-answer technique is of high validity in comparison with MCQs and true-false (9, 1 0). The same result was obtained in the present study.

It must be mentioned that to answer the short-answer questions was relatively difficult, and due to not having negative score, it contained the most number of wrong answers, but what is noticeable is that number of wrong answers in true-false questions was 26%, whereas in MCQs it was 32%. On the contrary, number of correct answers in true-false questions were the highest (65%).

Wass et al conducted a study on maximizing reliability of undergraduate clinical examinations, through combining, several question techniques given to 214 medical students in an 8 hours exam. He reported that 85.3% of them, in true-false method, gained the highest score (11), which supports our results

If in analysing the above mentioned result, one considers the cause as chance taking, it has to be pointed out that the highest number of unanswered questions (9%) were seen in that same technique.

The quality of the designed questions regarding difficulty and facility indices in different periods of basic sciences, physiopathology, and clinical sciences, did not reveal a common trend. In clinical sciences, half of the questions (45/90) were easy (F1>70%), and most of them were in the form of true-false (22/90), (Table2). Furthermore, the most number of questions were with discrimination index less than 0.3 (20.5%), (Fig1).

Thus in designing questions for the clinical period, the above mentioned points have to be taken into consideration.

It appears that the low number of correct answers in multiple-choice questions (60%) is possibly related to students facing wrong alternatives much more frequently.

Besides, the number of unanswered questions in this method is equal to that of the short-answer (8%), though the students knew that in the latter method there was no negative score for incorrect answers. Thus, it can be concluded that in multiple-choice questions, in spite of having negative scores, the higher frequency of wrong alternatives does not prevent the students from answering questions. They choose by chance and consider one of the choices as the correct one, without even knowing the correct answer.

The short-answer method, as compared with two other techniques, is the most valid one in assessing students' knowledge. So, the kappa test used to determine the common correct, incorrect, and unanswered questions in MCQs, and true-false, with short-answer method. In comparing the degree of agreement of each of the other two methods with the short-answer method, both the kappa coefficient and the coefficient of contingency are higher for the multiple-choice method than the true-false method. However, the kappa coefficients have very little difference with each other and are both in the "fair" range of agreement. Since the percentage of common correct answers are similar (78%), these differences are not due to correct answers, which are usually considered an index of students' knowledge, but differences in common incorrect (51% vs 39%) and common blank answers (28% vs 33%) were involved in obtaining this result (Table4).

Using kappa test, also showed that MCQs and true-false had the highest coefficient of contingency (0.402). Wass et al study also shows a correlation of 0.56 between short-answer and MCQs, and 0.46 between short-answer and true-false (11).

In a survey, Duffield et al. have analyzed the 381 students' views about the purposes and fairness of assessment in Newcastle Medical School. They concluded that a clear majority of students (8 1 %) agreed that, on the whole, assessment at this medical school was fair. Data interpretation papers (comprising a combination of multiple true-false, one best answer and short-answers) were perceived to be the fairest assessment tool (12).

Conclusion

Considering the highest kappa coefficient between true-false and multiple-choice methods, and high validity of short-answer technique, to enhance the validity and fairness of student evaluation, combination of the three methods is recommended, not the MCQs alone. No need to say that, by doing so, the rate of students learning and recall can be evaluated, and also the negative aspects of multiple-choice questions would be lessened.

Acknowledgements

We wish to thank the Ministry of Health and Medical Education for donating the grant which made this research possible. We would also like to express our appreciation to Dr. Maryam-o- Sadat Hosseini, Dr. Hossein Elyassi, Dr. Parviz Pakzad, Dr. Alireza Salek Moghadani, Dr. Behrooz Shafaghi, Dr. Manouchehr Shirvani, Dr. Jafar Forghanizadeh, Dr. Kourosh Gharagozli, and Dr. Homayoon Homayoonfar for meticulously designing questions and including them in the official examinations, thus increasing the credibility of this research. Most importantly, we express our special thanks to Professor Henry Walton, University of Edinburgh and Past-president of World Federation for Medical Education, for his critical reading of the manuscript and his invaluable comments and suggestions.

References

1- Hammond EJ, Mclndoe AK, Sansome AJ, and Spargo PM. Multiple-choice examinations: adopting an evidence-based approach to exam technique. Anaesthesia 1998; 53(1 I): 1 105-8.

2- Mavis BE, Cole BL, Hoppe RB. A survey of student assessment in US medical schools: the balance of breadth versus fidelity: Teach Learn Med 2001; 13(2): 74-9.

3- Schultheis NM. Writing cognitive educational objectives and multiple-choice test questions. Am J Health Syst Pharm 1998; 15; 55(22): 2397-401

4-Specific Student Assessment Techniques, in Student Evaluation: Teacher Handbook [Online]. 1991 Dec [cited 2001 May 12]; Available from: http://www.sasked.gov.sk.caldoc/Uoli£,ylstudevall

5- Frohlich ED. Medical Qualifying Exaniination: Rypins' Questions & Answers For Basic Sciences Review, 4 th ed. Philadelphia: Lippincott Williams & WiMns; 2001: 4.

6- Rassaian N, Ghandehari NS, Nakhaei S. and Tajasob B. Attitude and academic performance of medical students in research-centered teaching method. Med J Islani Rep Irn 2000; 14(3): 253-60.

7-Rassaian N. Long-term memory and learning through the use of research- centered teaching method. J Med Edu 2001; 1(l): 38-42.

8- A critical review of student assessment options [Online]. 2000 [cited 2001 Aug 27]; Available from: http:\\www.nmu.edu/soa/review.html.

9- Cox KR, Ewan CE. The Medical Teacher, Churchill Livingstone; 1982: 193-218.

10- MacAleer S. Objective testing: Dent JA, Harden RM (editors), A Practical Guide for Medical Teachers. London: Churchill Livingstone; 2001: 314-25.

11- Wass V, McGibbon D, Van der Vleuten C. Composite undergraduate clinical examinations: How should the components be combined to maximize reliability? Med Edu 2001:35(4):326-30.

12- Duffield KE, Spencer JA. A survey of medical students' views about the purposes and fairness of assessment. Med Edu 2002; 36(9): 879-86.