THE USE OF MULTIPLE CHOICE QUESTIONS IN MEDICAL EXAMINATION: AN EVALUATION OF SCORING AND ANALYSIS OF RESULTS

T.F. Ho
W.C.L. Yip
J.S.H. Tay

SYNOPSIS

Multiple choice questions (MCQ's) have become increasingly popular in both undergraduate as well as postgraduate medical examinations because of the claim of objectivity and reliability. To use this educational tool effectively, it is desirable that medical teachers be cognisant with the methodology. This paper makes use of the results of a class test consisting of 50 MCQ's to illustrate the principles and methods of grading and peer-referenced scoring. The individual questions are further analysed by calculating the difficulty index and the discrimination index. Critical evaluation of the individual questions together with analysis of these two indices would enable selection of suitable questions for re-use as well as identification of deficiency and ambiguity in teaching.

INTRODUCTION

The educational objectives in medicine as well as in other disciplines, are generally alloted to three 'domains' - cognitive, psychomotor and affective (1). Hence, medical examination should be designed to answer whether an undergraduate has achieved the above educational objectives by answering the following three questions: (1) What does he know (cognitive)? (2) What can he do (psychomotor)? and (3) What sort of person is he (affective)? Regretably the current medical examination system still could not answer these questions faithfully.

The structure of medical examination for undergraduates in Singapore patterns that of the United Kingdom which took shape towards the end of the nineteenth century. Traditionally it consists of three parts - the written papers, clinical examination and viva voce. Though the main format remains essentially unaltered, many changes have since been introduced, mainly to make the examination more objective and reliable.

Multiple choice questions (MCQ's) have recently become popular in medical examinations both for undergraduate and postgraduates because of the claim of objectivity and reliability in testing of the cognitive domain. However its almost idolatrous acclaim and its extensive use in medical examinations have lately been questioned by some experienced examiners (2). As the various pre- and para-clinical, as well as clinical departments of our Medical Faculty are adopting MCQ's in class tests, and as in some departments also in professional examinations, it is pertinent that medical teachers should be cognisant with the methodology of MCQ's.

This paper makes use of the results of a MCQ class test in physiology to illustrate the principles and methods of scoring and analysis of results. It is hoped that with better understanding of the methodology, MCQ's could be used more effectively to help fulfil the educational objectives.

Department of Physiology
Faculty of Medicine
National University of Singapore
College Road
Singapore 0316

T.F. Ho, MBBS, M.Med(Anaes.)
Lecturer

University Department of Paediatrics
National University of Singapore
Singapore General Hospital
Singapore 0316

W.C.L. Yip, MBBS, M.Med(Paed.), MRCP, DCH
Lecturer

J.S.H. Tay, MBBS, M.Med(Paed.), FRACP, B.D., M.D., AM
Assoc. Professor

MATERIALS AND METHODS

Fifty MCQ's on physiology of central nervous system were set to test a group of 60 University students, consisting of 37 first year dental students and 23 second year pharmacy students. Forty questions are of the true-false format and the remaining 10 belong to the 'multiple completion' variety or type K of the Hubbard and Clemen's classification (3). The test was given at the end of a course of 10 lectures on elementary physiology of central nervous system to these groups of students.

Each correct answer is awarded two marks and zero score is given to each wrong answer. There is no deduction of marks for wrong answer. The highest possible score is 100 and the lowest, 0.

The mean and standard deviation of the original scores were computed by standard statistical methods. The distribution of the scores was then determined whether it significantly differed from the normal Gaussian distribution by chi-square test for normal distribution (4) and by calculation of the skewness and kurtosis using the method of Fisher (5). For comparison, the original scores were converted to standard scores (Z and T scores). The students were then graded according to the original scores and to the standard scores into six grades, viz. distinction, A, B, C, D and E.

In the analysis of individual questions, the difficulty index and discrimination index were determined as follows. All the 60 students were ranked in order of merit from the highest score of 94 to the lowest score of 60. According to Ebel, R.L. (6), the first 27% of the students constitute the high group (H) and the last 27%, the low group (L). The difficulty index and discrimination index were then calculated according to Guilbert, J-J (1).

$$\text{Difficulty index} = \frac{H + L}{N} \times 100$$

$$\text{Discrimination index} = 2 \times \frac{(H - L)}{N}$$

where H = no. of correct answers in the high group
L = no. of correct answers in the low group
N = total no. of students in these two groups

The 50 MCQ's were then further analysed according to these two indices.

All statistical calculations were speedily executed by the newly installed Apple II microcomputer in the University Department of Paediatrics.

RESULTS

The arithmatic mean and standard deviation of the original scores of the dental and pharmacy students are tabulated in Table 1. As the difference between the two means is not statistically significant (P = 1.60), the original scores of the two groups of students are collected together for further analysis. The overall mean is 81.6 with a standard deviation of 7.6.

The distribution of the original scores of all the students is shown in Table 2 and Figure 1 is the graphic presentation. Table 3 shows that, by Chi-square test, the distribution of the original scorer does not differ significantly from the normal distribution (X^2 = 3.661, df = 3, $p > 0.1$). In addition, although the distribution shows slight tendency of negative skewness and kurtosis, statistical analysis (Figure 1) shows that the distribution of the scores is not significantly different from the normal Gaussian curve ($p > 0.05$).

Figure 2 shows the normal Gaussian distribution and conversion scale to convert the original scores to standard scores (Z scores and T scores). For example, the mean of the original scores is 82 and this corresponds to a standard Z score of 0 and a standard T score of 50. On the other hand, the corresponding Z and T scores for the original score of 74 are −1.0 and 40 respectively.

Table 4 and 5 show the results of grading according to original scores and standard T scores respectively. Note the dramatic change in the results of grading by the use of different scoring systems.

Table 6 shows the difficulty index and discrimination index of the 50 MCQ's. Table 7 and 8 display the classification of the 50 MCQ's according to the difficulty index and discrimination index respectively. The criteria are modified from Guilbert, J-J (1). Table 9 shows that easy (difficulty index ⩾70) and relatively easy (difficulty index 30 - 70) questions account for the great majority (98%) of the 35 MCQ's of low discrimination value (discrimination index ⩽0.24). On the other hand, there is only one difficult question (difficulty index <30) among these 35 MCQ's. Note that of the 15 MCQ's with good or excellent discrimination value (Table 9), almost half are easy questions, while the other half are average questions (difficulty index varying from 30 to 70).

Table 1 The means and standard deviations of the original scores

STUDENTS	NO.	Mean	Standard Deviation	Difference between the means

Dental	37	81.2	6.8	t = 0.53
				df = 58

Pharmacy	23	82.3	8.8	p = 1.60
Total	60	81.6	7.6	

t = student's t value
df = degree of freedom
p = 1.60, not significant

Table 2 The distribution of the Original scores

Original Scores	No. of students	% of students
58 - 61	1	1.7
62 - 65	1	1.7
66 - 69	1	1.7
70 - 73	4	6.7
74 - 77	11	18.3
78 - 81	8	13.3
82 - 85	11	18.3
86 - 89	12	20.0
90 - 93	10	16.7
94 - 97	1	1.7
Total	60	100.0

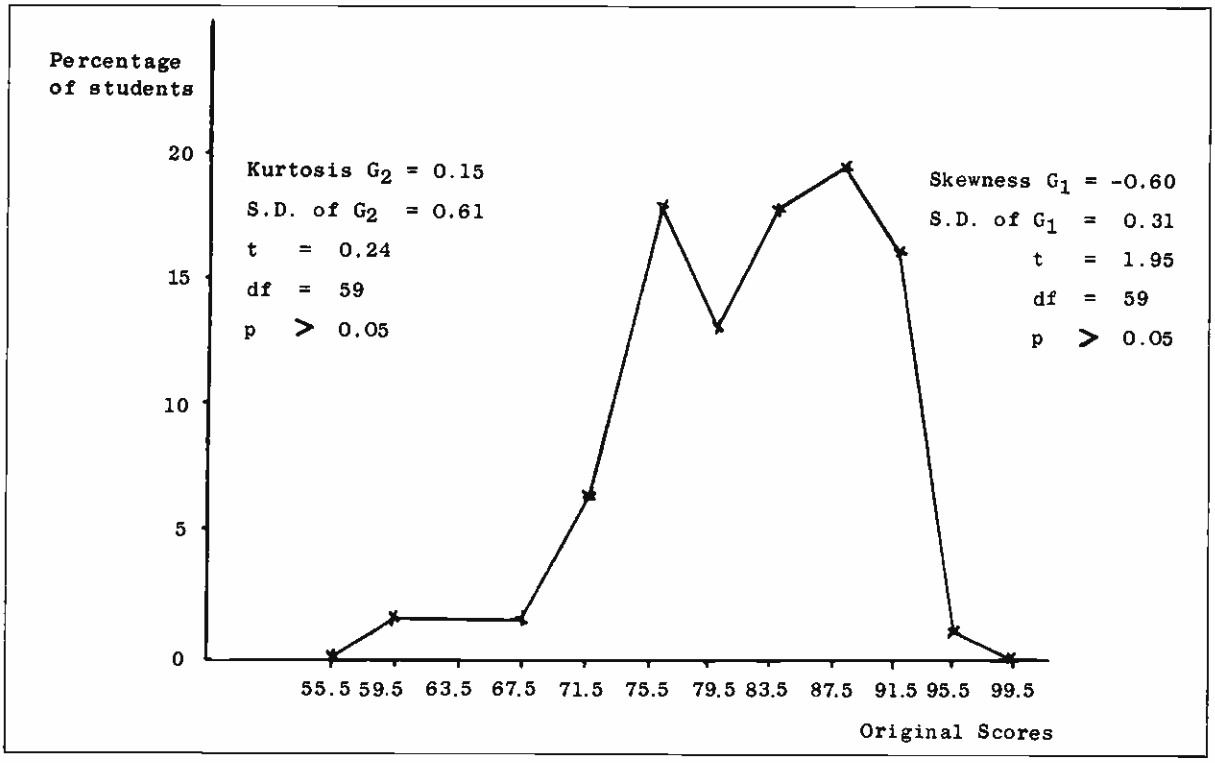

Fig. 1 : Frequency distribution of the original scores

Table 3 Chi-square test for normal distribution

Original Scores	Observed Frequency (O)	Expected Frequency (E)	$\frac{(O - E)^2}{E}$
<74	7	5.51	0.403
74 - 77	11	7.21	1.992
78 - 81	8	10.67	0.668
82 - 85	11	12.54	0.189
86 - 89	12	11.02	0.087
>90	11	13.05	0.322
Total	60	60.00	X^2 = 3.661 (df = 3, p>0.10)

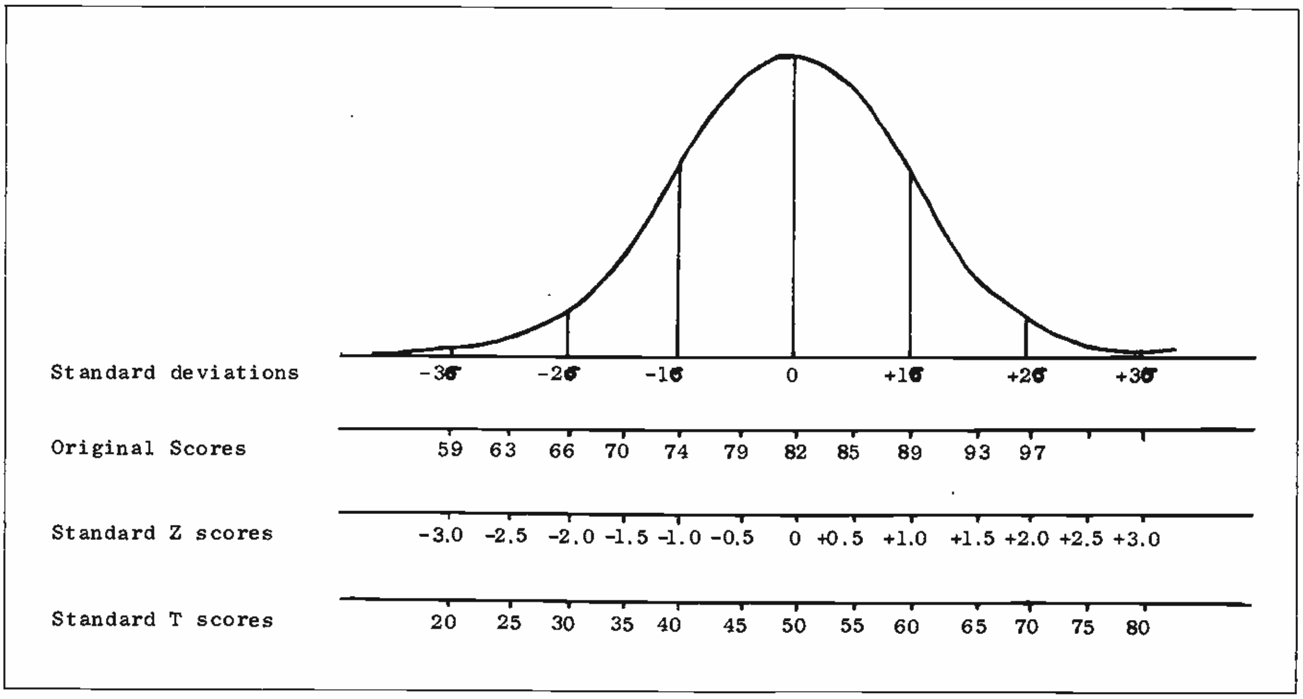

Fig. 2 : Gaussian distribution and conversion chart

Table 4 Grading according to original scores

Grading	Original Score	No. of students	% of students
Distinction	> 75	49	81.67
A	70 - 75	8	13.33
B	65 - 69	1	1.67
C	55 - 54	2	3.33
D	50 - 54	0	0
E	< 50	0	0
	Total	60	100.00

Table 5 Grading according to standard T Score

Grading	Standard T Score	No. of students	% of students
Distinction	⩾70	0	0
A	60 - 69	11	18.33
B	50 - 59	23	38.33
C	40 - 49	19	31.67
D	30 - 39	5	8.33
E	< 30	2	3.33
	Total	60	100.00

Table 6 The difficulty index and discrimination index of the 50 MCQ's

Question No.	Difficulty Index	Discrimination Index
1	94	0.13
2	72	0.19
3	82	0.38
4	67	0
5	100	0
6	100	0
7	94	0.13
8	94	0
9	100	0
10	88	0.13
11	59	0.56
12	81	0.13
13	97	0.06
14	34	0.31
15	81	0.13
16	94	0.13
17	78	0.31
18	97	0.06
19	78	0.44
20	91	0.06
21	97	0.06
22	100	0
23	72	0.44
24	84	0.19
25	100	0
26	97	0.06
27	97	- 0.06
28	94	0.13
29	97	0.06
30	69	0.38
31	34	0.44
32	94	0.13
33	78	0.44
34	91	0.19
35	100	0
36	88	0.25
37	66	0.44
38	38	0.13
39	97	0.06
40	97	0.06
41	66	0.19
42	94	0
43	47	0.44
44	63	0.50
45	17	0.19
46	97	0.06
47	72	0.56
48	84	0.19
49	63	0.38
50	88	0.13

Table 7 The distribution of the difficulty index of the 50 MCQ's

Difficulty Index	Interpretation*	Question No.	No. of Questions	% of Questions
>70	too easy	1-3, 5-10, 12-13, 15-29, 32-36, 39-40, 42, 46-48, 50	38	76
30 - 70	average, recommended	4, 11, 14, 30, 31, 37, 38, 41, 43, 44, 49	11	22
<30	too difficult	45	1	2
		Total	50	100

*Criteria as modified from Guilbert J-J.[1]

Table 8 The distribution of the discrimination index of the 50 MCQ's

Discrimination Index	Interpretation*	Question No.	No. of Questions	% of Questions
⩾0.35	Excellent discrimination	3, 11, 19, 23, 30, 31, 33, 37, 43, 44, 47, 49	12	24
0.25 - 0.34	Good discrimination	14, 17, 36	3	6
0.15 - 0.24	Marginal discrimination	2, 24, 34, 41, 45, 48	6	12
<0.15	Poor discrimination	1, 4-10, 12-13, 15-16, 18, 20-22, 25-29, 32, 35, 38-40, 42, 46, 50	29	58
		Total	50	100

*Criteria as modified from Guilbert J-J.[1]

Table 9 MCQ's of low discrimination value (discrimination index ⩽0.24)

Difficulty Index	No. of Questions	% of Questions
⩾70	31	89
30 - 70	3	9
<30	1	2
Total	35	100

Table 10 MCQ's of good or excellent discrimination value (discrimination index 0.25)

Difficulty Index	No. of Questions	% of Questions
⩾70	7	47
30 - 70	8	53
<30	0	0
Total	15	100

DISCUSSION

While all medical teachers agree that MCQ's are easy to mark, all would also appreciate that there are numerous problems, starting from the formulation of questions to the final analysis of results, in a MCQ test. As it is frequently impossible to predict the result of a set of new MCQ's attempted by a group of students, assessment of the marks of the students or scoring is not uncommonly beset with much difficulty. The current standard of grading in most of the departments in the Faculty of Medicine is based on the original score of the test according to the following scales:

Distinction	=	75
A	=	70 - 75
B	=	65 - 69
C	=	55 - 64
D	=	50 - 54
E	=	50

Hence it is not unusual in many class tests, or even in professional examinations when MCQ's are used, that majority of the students may obtain an 'E' grading if the questions are too difficult, or distinction, if too easy. However, this state of affair can be, and in fact should be prevented, especially in professional examinations, by grading according to peer-referenced scoring system (7), using standard T scores or Z scores rather than original scores. The pre-requisites of this conversion from original scores to standard scores are two. Firstly, the distribution of the original scores should not differ significantly from the Gaussian distribution. This is easily ascertained by chi-square test for normal distribution (Table 3), and according to Fisher, R.A. (5), by calculating the skewness (Gs1) and kurtosis (G_2) values. When both are not significantly different from zero, than the distribution would conform to the Gaussian curve (Figure 1). Secondly, having ascertained the above, with the knowledge of arithmatic mean and standard deviation, the original scores could then be easily placed against the standard T and Z scores, according to the conversion chart (Figure 2).

With reference to Table 4 and 5, it is obvious that instead of awarding 95% of the students a grading of A or distinction by the original scores, using the peer-referenced scoring system (7), grading according to standard T scores produces a more 'balanced' result. Hence the embarrassment of awarding too many candidates with distinction, or failing too many students because of too difficult or inappropriate MCQ's in professional examinations, can be avoided by this marking system.

Hitherto we have discussed only the scoring system and the result of the whole test. We would spend some time now looking at the individual questions. There are two important questions that always come to the mind of an examiner when he sets a MCQ. Firstly, is the MCQ too difficult, too easy or just about right? The second question is closely related to the first, and that is whether the MCQ could differentiate 'good' students from 'poor' students. Obviously questions which are too difficult or too easy have poor discrimination value. It is difficult and very often impossible to know the answers of these two questions before the test is administered to a group of students. Hence it is important for medical teachers to find out by calculating the difficulty and discrimination indices of all the MCQ's after marking the test paper.

According to Guilbert, J-J (1), MCQ's with difficulty index ranging from 30 to 70 are recommended for future use because these questions are likely to have satisfactory discrimination value. On the other hand, MCQ's with difficulty index outside the range are not recommended because they are either too easy (difficult index 70) or too difficult (difficulty index 30). However, it is important to realise that easy questions need not be useless questions, although they are likely to be less discriminative. For example, 76% of the 50 MCQ's on physiology of central nervous system have difficulty index over 70 and are therefore easy questions. This is not surprising because in a class test of this nature, the main educational objective is to find out at the end of the series of lectures whether majority of the students are able to understand the basic principles and to recall some essential facts. Hence it is expected that majority of the students will obtain correct answer for most of the questions. In addition, as is shown in Table 10, about half of the MCQ's with good or excellent discrimination indices are actually easy questions.

Guilbert, J-J (1) maintains that a discrimination index above 0.35 is excellent and that which ranges from 0.25 to 0.34 is good. MCQ's with index varying from 0.15 to 0.24 are considered to have marginal discrimination, while those less than 0.15, poor. Our analysis shows that only 30% of the MCQ's have good or excellent discrimination value (Table 9). The remaining 35 MCQ's have low discrimination index of less than 0.24. This is because majority of these are easy questions as explained above.

By analysing the difficulty index and discrimination index of a particular MCQ, we could evaluate the response of the students to that particular question so that we could ascertain not only whether that MCQ is too difficult, too easy or just about right for that group of students, but also whether it could differentiate 'good' students from 'poor' students. In addition, knowledge of these two indices will enable us to select suitable questions to be stored in the MCQ bank for future use, and also to detect flaws in the intrinsic structure of the questions which should then be discarded or suitably modified before re-use. Finally, critical evaluation of the results of MCQ's would also enable the teachers to detect deficiency and ambiguity in teaching as well as misconception of the students acquired from textbook.

REFERENCES

1. Guilbert J-J : Educational handbook for health personnel. WHO Offset Publication No. 35. Geneva: World Health Organisation, 1977.
2. Dudley HAF : Multiple choice tests : time for a second look? Lancet 1973; 2 : 195-196.
3. Hubbard JP and Clemens WV : Multiple-choice Examinations in Medicine. Philadelphia : Lea and Fabiger, 1961.
4. Loveday R : Statistics, second edition. London : Cambridge University Press, 1969.
5. Fisher RA : Statistical methods for research workers, forteenth edition. Edinburgh: Oliver & Boyd, 1970.
6. Ebel RL : Measuring education achievement. New Jersey USA: Prentice Hall, 1965.
7. Fleming PR, Sauderson PH, Stokes JF and Walton HJ : Examinations in medicine. Edinburgh, London and New York : Churchill Livingstone, 1976.