Comparative evaluation of multiple choice question formats

Introducing a knowledge score

Sheikh I. Rahim, *PhD, FRCPsych,* **Mahdi S. Abumadini,** *DPM, BCPsych.*

ABSTRACT

Objective: Over the last 6 years different multiple choice question (MCQ) formats have been used in postgraduate examinations for trainees in psychiatry. In phase 1 - K-type True/False (T/F) items with negative marking; in phase 2 combined T/F and type-A one-best answer (OBA) questions without negative marking; in phase 3 exclusively OBA without negative marking. The study compares the gross scores (GS) obtained with different MCQ formats, and introduces knowledge score (KS).

Methods: The study was conducted in the Saudi Council for Health Specialties, Riyadh, Kingdom of Saudi Arabia from 1996 to 2002. The mean percentile scores obtained by all postgraduate trainees sitting any Part I or Part II Saudi Board Examination in Psychiatry were subjected to a comparative analysis.

Results: A total of 110 candidates sat 18 examinations returning 143 papers containing a total of 32,375 MCQ options. Phase 1 generated lowest overall mean GS (47.8%), phase 3 occupied an intermediate position (53.1%) and phase 2 produced the highest score (68.3%). The KS, to the contrary, generated strikingly similar results for all the 3 phases (47.8, 50.5 and 49.5%) indicating that the marked differences in the GS were probably related to benefits obtained from guessing in the absence of negative marking. In this respect, the OBA produced considerably higher KS scores than the T/F, presumably due to its facilitating extra benefits from cueing, partial knowledge and judgement.

Conclusions: Different MCQ formats generate dissimilar quantitative results. The OBA format seems superior to the T/F format in crediting judgement and application of knowledge. In non-negatively marked MCQ tests, the suggested KS provides results comparable to those of negatively marked tests. Pass marks in MCQ tests should be calibrated according to the used format.

Neurosciences 2003; Vol. 8 (3): 156-160

Multiple choice questions (MCQs) grew out of a need for objective methods of assessing and ranking candidates.[1] They were considered superior to traditional essay questions for their simplicity, reliability, internal consistency and ability to cover a wide range of factual material.[2-7] Several formats have been used; the most popular of them are the true/false (T/F) and one-best-answer (OBA) questions, with[6,8-12] or without[2,13-15] negative marking for incorrect responses. The marks obtained by candidates depend not only on their actual knowledge,[2,4,16] but also on such diverse factors as the presence or absence of negative marking,[6,8,10] the precision of question formulations,[17-19] and the potential benefits from cueing,[20-22] partial knowledge,[8,11,23] and 'wild guessing'.[10,11,24] The pass marks in MCQ tests need regular calibration depending

From the Department of Psychiatry, King Fahd Hospital of the University, College of Medicine, King Faisal University, Al-Khobar, *Kingdom of Saudi Arabia*.

Received 29th September 2002. Accepted for publication in final form 29th December 2002.

Address correspondence and reprint request to: Dr. S. Rahim, Professor and Consultant Psychiatrist, Department of Psychiatry, King Fahd Hospital of the University, College of Medicine, King Faisal University, PO Box 10401, Al-Khobar 31952, *Kingdom of Saudi Arabia*. Tel. +966 (3) 8957911. Fax. +966 (3) 8993996. E-mail: sheikhidris35@hotmail.com

on the used format on the one hand, and on the relative difficulty of individual examinations on the other.[11,24-26] In the Kingdom of Saudi Arabia (KSA), the Saudi Council for Health Specialties (SCHS) adopts MCQs as its standard method of assessing postgraduate medical trainees in both Part I (in the middle of the 4-year training program) and Part II (final certifying) examinations.[27] Over the past 6 years the Psychiatry Program in SCHS employed MCQs in differing formats. In phase 1 (1997) all questions were of the K-type T/F format with negative marking for incorrect answers. In phase 2 (1996-1999) each paper was composed of K-type T/F and A-type OBA sections, both without negative marking. In phase 3 (2000-2001) all questions were of the A-type OBA and without negative marking too.

The aim of the present study is to compare the results of these examinations. By the rule of probability, we hypothesized that: 1. In phase 1, negative marking precluded significant gains from guessing and lead to lowest scores. 2. In phase 2, where no negative marking had been imposed, the T/F section yielded higher scores than the OBA section as it had 2 and a half times the probability of gain from guessing. 3. In phase 3, the exclusively OBA format yielded lower scores than those of phase 2. 4. OBA options provided better chances of utilizing cues and partial knowledge than the mutually independent T/F items, and 5. The end result is likely to be that candidates scored highest marks in phase 2, and least marks in phase 1.

Methods.

The study was conducted in the SCHS in Riyadh, KSA, from 1996 to 2002. Its material included the used MCQ papers, the answer sheets, the computer-generated scores, and the finally endorsed pass marks and pass rates. Over the 6-year study period a total of 110 candidates sat for 18 papers in 13 examinations (8 papers from 8 Part I and 10 papers from 5 Part II examinations). A total of 143 answered papers have been returned. In phase 1, 2 candidates returned 4 papers offering 1,200 T/F response options. In phase 2, 55 candidates returned 68 papers offering 17,000 T/F and 4,725 OBA options. In phase 3, 53 candidates returned 71 papers offering 10,650 OBA options. The overall study material consisted of 33,575 response options.

Operational definitions. 1. The gross score (GS) is the officially accepted computer-generated percentile mark given to the examinee, regardless of how much of it has been earned by actual knowledge. 2. The pass mark (PM) is the minimum percentile GS that enabled any candidate to pass the specified examination. 3. The pass rate (PR) is the percentile proportion of candidates who had passed the specified examination. 4. The knowledge score (KS) is a hypothetical, probability estimate of the amount of actual knowledge needed by the candidate to obtain on average his achieved GS. 5. The confounding score (CS) is a complementary construct representing that portion of the GS which was obtainable independently of knowledge.

Measurement. The GS, PM and PR were retrieved from official records of the SCHS. The KS and CS were computed from the GS according to the rule of the probability as follows. In phase 1, the KS was considered equal to the GS on the assumption that, in large populations of dichotomous, mutually independent T/F responses, negative marking would, by the rule of probability, neutralize the effects of wild guessing, and the benefits from partial knowledge would be proportionate to the degree of that knowledge. In the 2 other phases, where no penalty had been imposed on incorrect responses, and where candidates routinely attempted virtually all questions irrespective of their degree of conviction about the correctness of their responses, the KS was considered equal to the GS minus that part of it, which the candidate needed no knowledge to obtain by mere chance probability (CS). This means that for the T/F the GS would be equal to the actually known answers (the KS) plus 50% of the guessed ones (the CS); and for the OBA to be equal to the KS plus 20% of the guessed answers. Thus, for T/F the KS would be twice the GS minus 100, and for OBA would be 1.25 the GS minus 25.

Data entry and analysis was carried out using the Statistical Package for Social Sciences for Windows version 10.01. Scores were compared by means and their standard deviations and were tested by F-value. The correlation between the T/F and OBA scores in phase 2 was tested by their correlation coefficient. Only statistically significant data ($p < 0.05$) is discussed.

Results.

Table 1 provides the mean values of GS for the T/F, OBA, the total, the pass mark and pass rate for each examination. The overall mean percentile GS score was 64.3, varying remarkably from 47.8 in phase 1, where the examination was exclusively of the T/F format with negative marking, to 53.1 in phase 3, where it was exclusively of the OBA format without negative marking, to 68.3 in phase 2, where it was combined T/F - OBA formats without negative marking. Within phase 2, candidates scored significantly higher percentile scores in the T/F (72.2) than in the OBA sections (64.5). This tendency was consistently observed at the level of individual examinations. It reflected itself on the officially approved pass marks, which varied from 47.2 in phase 1 to 58.0 in phase 3 and 62.2 in phase 2. Notably, the default pass mark of 70% as stipulated by the Examination Regulations,[27] was applied in only 2 examinations, both of which were from phase 2. The overall PR was 72% of the examinees, being significantly higher in phase 2 (76.5%) than in phase 3 (66.2%), and in Part I (78.8%) than in Part 2 examinations (66.2%). The correlation between T/F and OBA scores (applicable to phase 2 only) was moderately positive (r=0.55). In phase 1, the candidates were collectively offered a total of 1,200 T/F items with

Table 1 - Mean percentile gross scores in MCQ questionnaires.

DATE	PART	PAPERS	CANDID	T/F*	OBA*	TOTAL	PASS MARK	PASS RATE
Phase 1								
Nov 1997	2	4	2	47.8	-	**47.8**	47.2	100.0
Phase 2								
June 1996	1	10	10	70.8*	55.0	**62.9**	62.2	70.0
Nov 1997	1	5	5	78.4*	65.9	**72.2**	70.0	80.0
May 1998	1	10	10	72.1	70.8	**71.4**	73.0	70.0
Nov 1998	1	3	3	68.1*	56.1	**61.9**	63.6	66.7
Nov 1998	2	16	8	74.8*	67.3	**71.0**	66.9	87.5
Sept 1999	1	14	14	70.5	65.6	**68.0**	62.4	71.4
Oct 1999	2	10	5	70.2	63.6	**66.9**	65.5	80.0
Phase 3								
Sept 2000	1	15	15	-	64.9	**64.9**	60.0	80.0
Oct 2000	2	20	10	-	60.3	**60.3**	61.4	70.0
May 2001	1	5	5	-	50.4	**50.4**	63.3	20.0
Sept 2001	1	15	15	-	61.2	**61.2**	60.0	53.3
Nov 2001	2	16	8	-	58.6	**58.6**	58.4	75.0
Totals by Phase								
Phase 1	2	4	2	47.8	-	**47.8**	47.2	100.0
Phase 2	1,2	68	55	72.2*	64.5	**68.3**	62.2	76.5
Phase 3	1,2	71	53	-	53.1	**53.1**	50.4	66.2
Totals by Part								
Part 1	1	77	77	71.7*	58.8	**64.8**	60.0	66.2
Part 2	2	66	33	73.0*	65.5	**63.7**	58.0	78.8
GRAND TOTAL	**1,2**	**143**	**110**	**72.2***	**62.4**	**64.3**	**58.0**	**72.0**

Candid - number of candidates sitting the examination(s), T/F - K-type True/False Format, OBA - A-type One-Best-Answer Format, *statistically significant difference between T/F and OBA, MCQ - multiple choice question

Table 2 - Mean percentile knowledge score in MCQ examination(s).

DATE	PART	PAPERS	CANDID	T/F*	OBA*	TOTAL	CS	CS%
Phase 1								
Nov 1997	2	4	2	47.8	-	**47.8**	-	-
Phase 2								
June 1996	1	10	10	41.7	43.8	**42.7**	20.2	33.1
Nov 1997	1	5	5	56.8	57.4	**57.1**	15.1	21.1
May 1998	1	10	10	44.2	63.5*	**53.8**	17.6	25.6
Nov 1998	1	3	3	36.1	45.2	**40.7**	21.3	34.7
Nov 1998	2	16	8	49.5	59.1*	**54.3**	16.7	23.9
Sept 1999	1	14	14	40.9	57.0*	**49.0**	19.1	29.3
Oct 1999	2	5	5	40.4	54.4*	**47.4**	19.5	30.1
Phase 3								
Sept 2000	1	15	15	-	56.2	**56.2**	8.8	14.2
Oct 2000	2	20	10	-	50.4	**50.4**	9.9	17.3
May 2001	1	5	5	-	38.0	**38.0**	12.4	25.8
Sept 2001	1	15	15	-	51.0	**51.5**	9.7	16.4
Nov 2001	2	16	8	-	48.2	**48.2**	10.4	18.8
Totals by Phase								
Phase 1	2	4	2	47.8	-	**47.8**	-	-
Phase 2	1,2	68	55	43.4	55.6*	**49.5**	18.3*	27.8*
Phase 3	1,2	71	53	-	50.5	**50.5**	9.9	17.4
Totals by Part								
Part 1	1	77	77	43.4	53.2*	**49.0**	14.6	23.3
Part 2	2	66	33	46.0	52.7*	**51.6**	13.9	21.5
GRAND TOTAL	**1,2**	**143**	**110**	**44.4**	**53.0***	**50.0**	**14.3**	**22.3**

Candid - number of candidates sitting the examination(s), T/F - K-type True/False Format, OBA - A-type One-Best-Answer Format, *statistically significant difference, CS - confounding score, CS% - CS as a percent of the gross score, MCQ - multiple choice question negative marking. They declined to answer 199 (16.6%) items. They scored positive marks from 787 (65.6%) correct responses, and negative marks from 214 (17.8%) wrong responses. Their mean score came out to be 47.8%. It can easily be calculated that, had there been no negative marking and the candidates consequently attempted all items, they could have obtained an estimated probability score of 73.9%, a figure strikingly similar to that of the T/F without negative marking in phase 2.

Table 2 displays the estimated KS and CS. The overall mean KS was 50%, 14.3% marks below the corresponding figure for GS. The difference between the GS and KS (namely, the CS) was nearly three-fold bigger in the T/F (27.8%) than in the OBA tests (9.4%). Unlike the GS, the KS showed no significant variation across phases or individual examinations. While the GS was consistently higher in the T/F than in the OBA items, the KS was, to the contrary, significantly higher in the OBA than in the T/F items. The CS was twice as high in phase 2, where the examination was of mixed T/F-OBA format than in phase 3, where it was exclusively of the OBA format (18.3% and 9.9%). This CS constituted 39.4% of the GS for all the T/F tests compared with 17.2% of it for all the OBA tests. Within phase 2, where performance of the same candidate in the T/F and the OBA sections of the same examination could be compared, the CS was threefold higher in the T/F than in the OBA sections (27.8% and 8.9%).

Discussion.

Limitations. Firstly, the material is drawn from a limited number of examinations involving a relatively small number of examinees. However, the unit of analysis has not been the number of examinations or of candidates but the whole 33,575 responses to individual MCQ items. Secondly, though genuine efforts have been exerted by the SCHS[27] to exclude faulty or ambiguous question formulations,[17,18,28] some subtle flaws might have escaped notice. These limitations should be born in mind when evaluating the present data.

Findings. The basic observation is that different MCQ formats generated quantitatively dissimilar results. As expected, negative marking produced lowest scores. Its proponents wanted to discourage guessing by precluding its benefits;[8,29] its critics considered that 'unfortunate', arguing that, 'in much of medicine, informed and educated guessing are exactly what is needed'.[6] Concern has been expressed that, as the fear of losing marks inhibits most people from answering items they do know, but about which they do not feel fully confident, this might produce a divergence between candidates on the basis of their readiness to take risks, rather than on their knowledge.[6] Our data showed that in phase 1, where negative marking has been applied, the candidates refrained from responding to one third of the estimated number of items about which they had been uncertain. In the other phases, where no negative marking had been imposed, the candidates obtained considerably higher GS in the T/F than in the OBA items, but their KS were, to the contrary, higher in the OBA than in the T/F items. This indicates that the OBA format offers more chances of utilizing cueing,[20-22] partial knowledge,[8,11,24] and judgment.[2] Koeslag and Melzer[30] differentiated 3 types of guessing in MCQ tests: guessing due to partial ignorance, guessing due to total ignorance, and the so-called 'antiknowledge' which they defined as 'recording an incorrect response in the firm belief that it is the correct one'. Holden[8] pointed out that candidates, in reality, have a spectrum of certainty about the correctness of a particular response ranging from knowing with absolute conviction to knowing nothing at all (wild guessing). Rather than viewing this spectrum as a succession of discretely identifiable entities, we considered it a one-dimensional continuum of the degrees of knowledge on the issue in question ranging from 'zero knowledge' (absolutely blind guessing) to 100 percent knowledge (full certainty of the correct response). By the rule of the probability, the likelihood of making a correct response to a particular item would be, on average, proportionate to the position of the candidate's knowledge along that continuum. Our suggested KS not only eliminates gains from blind guessing, but also variably credits partial knowledge for whatever it is worth. The superiority of OBA on T/F in the KS of non-negatively marked examinations confirms reports on the greater chances of benefiting from cueing, partial knowledge and judgement in OBA than in T/F tests.[2,31] It should be emphasized that the KS has no effect on the rank order of candidates sitting the same examination. It is specially useful for comparing results of examinations conducted with different MCQ formats, and for calibration of equivalent pass marks. As a probability estimate, its reliability depends on the sample size of the studied observations. It is unsafe to use it for comparing individuals: luckier ones might get more than their estimated group mean at the expense of the unlucky ones who get less.

In conclusion, MCQs are a valuable method of testing factual knowledge. Different formats generate considerably different results. Negative marking produces lowest scores. In the absence of negative marking, the K-type T/F format yields higher gross scores, but lower KS than the A-type OBA format. The suggested KS is useful for comparing the results of different examination formats and in calibrating suitable pass marks, but it is unsafe in comparisons between individual candidates. The OBA seems superior to the T/F format in that it measures not only factual knowledge, but also its sensible application. Further, large-scale studies are needed to confirm these findings.

Acknowledgment. The authors are greatly indebted to HE Professor Hussain Al-Freihi, Secretary General of the Saudi Council for Health Specialties for approving this study, and for his valuable advice and comments. Due gratitude is extended to Professor Abdalla Al-Subaie, Chairman of the Scientific Council of Psychiatry, and to all the members of the Council without whose encouragement and support this work could not have been accomplished.

References

1. Anderson J. The Multiple choice Questions in Medicine. Tunbridge Wells (UK): Pitman Medicals; 1976.
2. Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences. 2nd ed. Philadelphia (PA): National Board of Medical Education; 1998.
3. Lowy FH, Prosen H. The Canadian certification examination in psychiatry. III. Towards better certification techniques. ***Can Psychiatr Assoc J*** 1979; 24: 292-301.
4. Norman GR, Smith EK, Powles AC, Rooney PJ, Henry NL, Dodd PE. Factors underlying performance on written tests of knowledge. ***Med Educ*** 1987; 21: 297-304.
5. Hill DA. Role of the pre-test in the progressive assessment of medical students. ***Aust N Z J Surg*** 1992; 62: 743-746.
6. Simpson MA. MCQ Tutor: Psychiatry. London (UK): William Heinemann Medical Books Ltd; 1983.
7. Norcini JJ, Swanson DB, Grosso LJ, Shea JA, Webster GD. A comparison of knowledge, synthesis and clinical judgment: multiple-choice questions in the assessment of physician competence. ***Evaluation & the Health Professions*** 1984; 7: 485-499.
8. Holden NL. The MRCPsych examination. MRCPsych supplement. ***Br J Hosp Med*** 1989: 42: 415-418.
9. Freeman C. MCQs for Psychiatric Studies. London (UK): Churchill Livingstone Publishers; 1988.
10. Fleming PR. The profitability of 'guessing' in multiple choice question papers. ***Med Educ*** 1988; 22: 509-513.
11. Marshall EJ. Multiple Choice Questions for the MRCPsych Part I Examination. Oxford (UK): Butterworth-Heinemann; 1995.
12. Harden RM, Brown RA, Biran LA, Dalles Ross WP, Wakeford RE. Multiple choice questions: to guess or not to guess. ***Med Educ*** 1976; 10: 27.
13. Mavis BE, Cole BL, Hoppe RB. A survey of student assessment in U.S. medical schools: the balance of breadth versus fidelity. ***Teaching & Learning in Medicine*** 2001; 13: 74-79.
14. Easterling WE Jr. In-training examinations for residents in obstetrics and gynecology 1975 to 1978. ***Am J Obstet Gynecol*** 1979; 133: 733-741.
15. Lynch TG, Woelfl NN, Steele DJ, Hanssen CS. Learning style influences student examination performance. ***Am J Surg*** 1998; 176: 62-66.
16. Robinowitz HK, Hojat M. A comparison of a modified essay question and multiple choice formats: their relationships to clinical performance. ***Fam Med*** 1989; 21: 364-367.
17. Holsgrove G, Elzubeir M. Imprecise terms in UK medical multiple-choice questions: what examiners think they mean. ***Med Educ*** 1998; 32: 343-350.
18. Vahalia KV, Subramaniam K, Marks SC Jr, De Souza EJ. The use of multiple-choice tests in anatomy: common pitfalls and how to avoid them. ***Clin Anat*** 1995; 8: 61-65.
19. Melzer CW, Schach SR, Koeslag JH. Misconceptions and miscarriages in multiple choice questions. ***S Afr Med J*** 1976; 50: 583-587.
20. Fajardo LL, Chan KM. Evaluation of medical students in radiology. Written testing using uncued multiple-choice questions. ***Invest Radiol*** 1993; 28: 964-968.
21. Veloski JJ, Rabinowitz HK, Robeson MR. A solution to the cueing effects of multiple choice questions: the Un-Q format. ***Med Educ*** 1993; 27: 371-375.
22. Anderson J. The MCQ controversy - a review. ***Med Teach*** 1981; 3: 150-156.
23. Hammond EJ, McIndoe AK, Sansome AJ, Spargo PM. Multiple choice examination: adopting an evidence-based approach to exam technique. ***Anaethesia*** 1999; 53: 1105-1108.
24. Howells R. Multiple Choice Questions for the MRCPsych Part II Clinical Topics Examination. Oxford (UK): Butterworth-Heinemann; 1995.
25. Morgan GH, Hill PD. MCQs in the MRCPsych examinations. ***Psychiatric Bulletin*** 1991; 15: 108.
26. Levi MI. MCQs for the MRCPsych Part II. London (UK): Kluwer Academic Publishers; 1993.
27. Saudi Council for Health Specialties. General Examination Rules and Regulations. Riyadh (KSA): SCHS; 2001.
28. Cox KR. How to construct a fair multiple choice paper. In: Cox KR, Ewan CE, editors. The Medical Teacher. Edinburgh (UK): Churchill Livingstone Publishers; 1982. p. 211-214.
29. Slade PD, Dewey ME. Role of grammatical clues in multiple choice questions: an empirical study. ***Med Teach*** 1983; 5: 146-148.
30. Koeslag JH, Melzer CW. The incorrect response in multiple-choice examinations. ***S Afr Med J*** 1981; 60: 591-592.
31. Epstein RM, Hundert EM. Defining and assessing professional competence. ***JAMA*** 2002; 287: 226-235.