Original Article (CPD)

Relationship Between Item Difficulty and Discrimination Indices in True/False-Type Multiple Choice Questions of a Para-clinical Multidisciplinary Paper

Si-Mui Sim,[1] *BSc, PhD*, Raja Isaiah Rasiah,[2]

Abstract

Introduction: This paper reports the relationship between the difficulty level and the discrimination power of true/false-type multiple-choice questions (MCQs) in a multidisciplinary paper for the para-clinical year of an undergraduate medical programme. Materials and Methods: MCQ items in papers taken from Year II Parts A, B and C examinations for Sessions 2001/02, and Part B examinations for 2002/03 and 2003/04, were analysed to obtain their difficulty indices and discrimination indices. Each paper consisted of 250 true/false items (50 questions of 5 items each) on topics drawn from different disciplines. The questions were first constructed and vetted by the individual departments before being submitted to a central committee, where the final selection of the MCQs was made, based purely on the academic judgement of the committee. Results: There was a wide distribution of item difficulty indices in all the MCQ papers analysed. Furthermore, the relationship between the difficulty index (P) and discrimination index (D) of the MCQ items in a paper was not linear, but more dome-shaped. Maximal discrimination (D = 51% to 71%) occurred with moderately easy/difficult items (P = 40% to 74%). On average, about 38% of the MCQ items in each paper were "very easy" (P ≥75%), while about 9% were "very difficult" (P <25%). About two-thirds of these very easy/difficult items had "very poor" or even negative discrimination (D ≤20%). Conclusions: MCQ items that demonstrate good discriminating potential tend to be moderately difficult items, and the moderately-to-very difficult items are more likely to show negative discrimination. There is a need to evaluate the effectiveness of our MCQ items.

Ann Acad Med Singapore 2006;35:67-71

Key words: Assessment methods, Educational measurement, Item analysis, Year II medical test

Introduction

Multiple-choice questions (MCQs) are used more and more in departmental examinations or as comprehensive examinations at the end of an academic session.[1] They may be used to determine progress or to make decisions regarding the certification of a candidate.[2] They may also be used to identify strengths and weaknesses in students as well as to provide feedback to teachers on their educational actions. The manner in which the test questions are prepared and put together to form an examination, and the procedure for scoring, analysing and reporting the results, all have a bearing upon the conclusions drawn from the performance of the individuals and groups tested.

MCQs, whether in the format of "true/false" or "one-best-answer", are expressly designed to assess knowledge. They have the advantage of sampling broad domains of knowledge efficiently and hence reliably.[3] This one characteristic of MCQs is sufficient to ensure that its edge in reliability more than compensates for some perceived failings in validity. Concerns have been voiced that most MCQs tend to measure factual recall and recognition of isolated facts. But if carefully constructed, MCQs (especially one-best-answer-type) may also test higher-order thinking skills.[4] Therefore, MCQs remain a useful assessment instrument, despite some limitations and objections.

Before 1998, the assessment methods used in the MBBS

[1] Department of Pharmacology
[2] Office of the Dean
Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia
Address for Reprints: Dr Debra Si-Mui Sim, Department of Pharmacology, Faculty of Medicine, 50603 Kuala Lumpur, Malaysia.
Email: debrasim@um.edu.my

This paper has been selected for inclusion in the Academy of Medicine's Continuing Professional Development online Distance Learning Programme and will be accessible after the next quarterly update in April 2006, via a secured Login. Please visit the Academy's website at http://www.ams.edu.sg for more details.

programme of the University of Malaya had traditionally included long essays, true/false-type MCQs and short-answer or "spot-test" practical examination in the para-clinical years. These test questions, which were discipline-based, were developed and vetted within the departments that taught the respective disciplines, and administered by the individual departments.

Since 1998, with the introduction of our New Integrated Curriculum (NIC), short-answer-type questions (SAQ) have replaced long essays, while true/false-type MCQs and short-answer or "spot-test" practical examination, now known as objective structured practical examination (OSPE), remain. However, all examination papers are now multidisciplinary, with some integration across the disciplines, such as in the scenario-orientated SAQ. All examination questions are now centrally vetted and the examination papers are administered by the Office of the Dean.

Although some basic form of item analysis of the MCQ tests has been carried out routinely since the beginning of the NIC, there has been no evidence that the data generated have been used to help develop or select subsequent MCQ items. So just how "good" are our MCQ tests? How effective are the individual MCQ items in predicting the students' overall performance in the whole MCQ test paper? Have we maintained similar standards of MCQ tests from year to year? These are some of the questions we attempted to answer when auditing the MCQ of selected examination papers.

The purpose of this study, therefore, was to examine the quality of our current Year II (para-clinical) multidisciplinary true/false-type MCQ tests, and to see if there was any relationship between the item difficulty index and the item discrimination index values in these MCQ tests.

Materials and Methods

Construction and Selection of MCQ Items

The MCQ items were first written by individual teachers and vetted at their respective departments for content accuracy. The vetted questions (newly written or extracted from the bank) were then chosen by the departmental coordinator and/or head before being submitted to the central vetting committee, which consisted of mostly senior academic staff representing each department concerned and was chaired by the Year II Coordinator. The final selection of the MCQ items for an examination paper was done by this central committee, and was based purely on the academic judgement and examination experience of the committee members present. After the final vetting by the central committee, the selected MCQ items were formatted by the Office of the Dean for the examination.

Data Collection

MCQ items taken from past Year II Parts A, B and C examinations were analysed for level of difficulty and power of discrimination. Each of these examinations was carried out at the end of a term that consisted of 12 to 16 weeks of teaching. We included in this study the MCQ papers from all the 3 (Parts A, B and C) examinations of 1 academic session (2001/2002), as well as the MCQ papers for Part B examination from 2 other consecutive sessions (i.e., 2002/2003 and 2003/2004). There were 155 students who sat for the examinations in session 2001/2002, 212 students in 2002/2003 and 214 students in 2003/2004. Each end-of-term examination covered different topics, grouped generally according to organ-systems and also included some foundational (core) topics. However, some degree of overlap in the topics tested between one examination and another occurred.

Scoring of MCQs

The MCQ paper contained 50 questions drawn from the 4 major para-clinical disciplines – Pathology, Medical Microbiology, Parasitology and Pharmacology – and could also include other disciplines such as Medical Statistics and Epidemiology, Neuroanatomy and Neurophysiology in some of the examinations. The MCQ paper formed part of a 3-hour written paper and was to be completed in 75 minutes. Each question consisted of a stem and 5 completing phrases, and students were required to categorise each of the 5 sentences thus constructed (the *items*) as True or False. A correct response to an item was awarded 1 mark, while an incorrect response would result in the deduction of 1 mark, and a no-attempt or blank response (indicating "I don't know") was given 0 marks. However, there was no carrying over of negative marks from one question to another. Thus, the maximum total score for any one question was 5 marks while the minimum total score was 0 (and not -5) marks.

Item Analysis

The results of students' performance in these MCQ tests were then used to determine the difficulty index and discrimination index of each MCQ item in the respective tests. In this study, the *item difficulty index (P)* refers to the percentage of the total number of correct responses to the test item. It is calculated by the formula $P = R/T$, where R is the number of correct responses and T is the total number of responses (i.e., correct + incorrect + blank responses). Hence, the higher this index value, the lower is the difficulty, and the greater the difficulty of an item, the lower is its index. The *item discrimination index (D)*, however, measures the difference between the percentage of students in the upper group (P_U), i.e., the top 27% scorers, who obtained the correct response, and the percentage of those in the lower group (P_L), i.e., the bottom 27% scorers, who obtained the correct response; thus $D = P_U - P_L$. The higher the discrimination index, the better the item can determine the difference, i.e., discriminate, between those students with high test scores and those with low ones.[5]

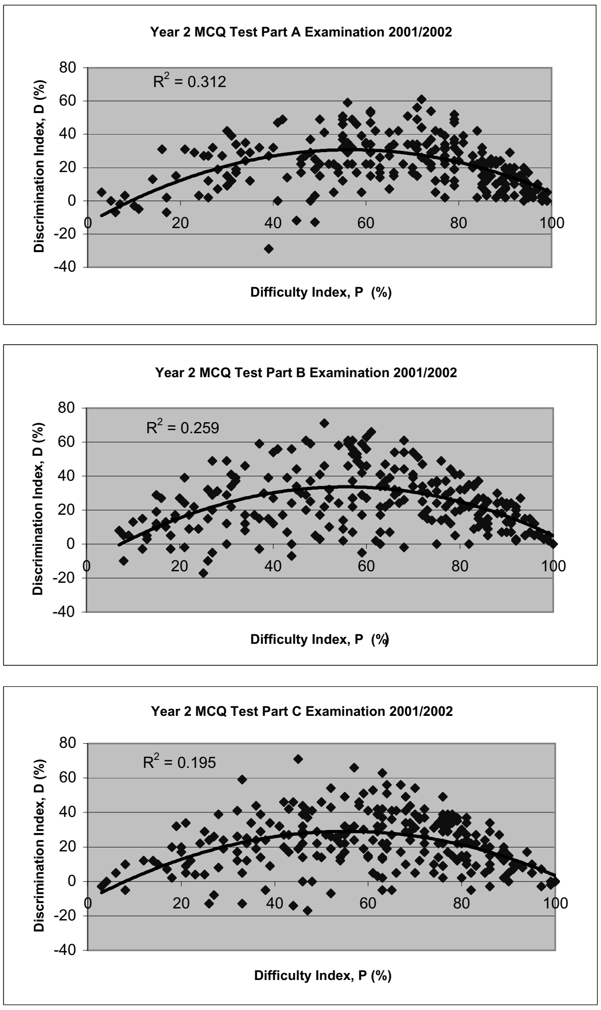

Fig. 1. The relationship between item difficulty index and discrimination index values of the MCQ papers (n = 250 test items) for Parts A, B and C examinations, administered to 155 Year II medical students in the University of Malaya, Session 2001/2002.

Statistical Analysis

All data are reported as mean ± SD of n items. The relationship between the item discrimination index and difficulty index values for each test paper was determined using curve estimation regression analysis [Statistical Package for the Social Sciences (SPSS) version 13; SPSS Inc., Chicago, Illinois, USA], and the coefficient of determination is given by R^2. A P value of <0.05 was considered to indicate statistical significance.

Results

Figure 1 shows the relationship curves between the discrimination index and difficulty index values for 3 of the 5 MCQ test papers analysed, which were for Session 2001/2002. Using curve estimation, it was found that the independent variable difficulty index contributed 31.2%, 25.9% and 19.5% (as shown by the R^2 values on the graphs, P <0.001) of the total variance of the outcome discrimination index for Parts A, B and C examinations, respectively. There seemed to be a gradual decrease in the R^2 values from Part A to Part C examination of the same academic session. The relationship between the two indices for Part B examinations of Sessions 2002/2003 and 2003/2004 (R^2 = 0.278 and 0.256, respectively; P <0.001) closely resembled that for Session 2001/2002 (R^2 = 0.259). Further analysis of the data indicated that there was a wide spectrum of level of difficulty among the MCQ items in all the papers. The difficulty index of these papers ranged from as low as 1% to 7% ("extremely difficult" items) to as high as 99% to 100% ("extremely easy" items), as shown in Table 1.

Regardless of the topics examined or sessions, the relationship between the difficulty index and discrimination index values of the MCQ items in a paper was not linear, but more dome-shaped. Initially, the discrimination power increased with the level of difficulty of the items, until it reached a plateau (discrimination index of 51% to 71%) with moderately easy/difficult items (difficulty index of 40% to 74%), and then began to decline with further increase in difficulty (difficulty index <25%).

Table 1. Mean Difficulty Index (P) and Discrimination Index (D) for Each MCQ Paper Analysed for Parts A, B and C Examinations (n = 250 test items)

Academic session	Part	No. of students	Difficulty index P (%)		Discrimination index D (%)	
			Mean ± SD	*Range*	Mean ± SD	*Range*
2001/2002	A	155	66.7 ± 23.8	*3 to 99*	33.3 ± 15.2	*-29 to 61*
2001/2002	B	155	59.6 ± 25.0	*7 to 100*	24.4 ± 17.4	*-17 to 71*
2001/2002	C	155	59.6 ± 22.6	*3 to 100*	22.2 ± 16.4	*-41 to 71*
2002/2003	B	212	60.4 ± 26.8	*1 to 100*	20.6 ± 16.1	*-26 to 58*
2003/2004	B	214	63.1 ± 24.7	*4 to 100*	21.9 ± 17.3	*-28 to 60*

Table 2. Proportion of "Very Easy" (P ≥75%) and "Very Difficult" (P <25%) Items for Each MCQ Paper Analysed for Parts A, B and C Examinations (n = 250 test items)

Academic session	Part	Topics covered in examination	"Very easy" items	"Very difficult" items
			% (no.)	% (no.)
2001/2002	A	Core topics for the different disciplines	45.6 (114)	6.0 (15)
2001/2002	B	CVS, Resp, GIT/Hep, Blood, Renal	33.2 (83)	11.6 (29)
2001/2002	C	Endo, Repr, CNS, MS, Infections, Skin	30.4 (76)	8.0 (20)
2002/2003	B	CVS, Resp, GIT/Hep, Blood, Renal	40.4 (101)	12.0 (30)
2003/2004	B	CVS, Resp, GIT/Hep, Blood, Renal	38.8 (97)	8.0 (20)
		Mean ± SD	37.8 ± 6.0	9.1 ± 2.6

CNS: central nervous system; CVS: cardiovascular system; Endo: endocrine; GIT/Hep: gastrointestinal tract and hepatobiliary; MS: musculoskeletal; Repr: reproductive; Resp: respiratory

On average, 37.8 ± 6.0% (mean ± SD) of the 250 MCQ items in each paper had a difficulty index of ≥75% ("very easy" items), while about 9.1 ± 2.6% items had a difficulty index of <25% ("very difficult" items), as shown in Table 2. About two-thirds of these "very easy" and "very difficult" items had poor or even negative discrimination (D ≤20%). Generally, discrimination correlated positively with difficulty at the "easy end" (P between 80% and 100%) of the curve, but negatively at the "difficult end" (P between 0% and 20%) of the curve.

Discussion

As with other health professional training, the effective measurement of knowledge is an important component of both medical education and practice.[6] Furthermore, the methods used to analyse the evidence resulting from the tasks (i.e., *interpretation*) need to be aligned with the aspects of achievement that are to be assessed (i.e., *cognition*) and the tasks used to collect evidence about students' achievement (i.e., *observation*).[7] Therefore, it is important for us to evaluate our MCQ items to see how effective they are in assessing the knowledge of our medical students in the para-clinical year of training, and in predicting their total test scores.

Many methods have been developed to calculate the discriminatory power of individual items; e.g., discrimination index, biserial correlation coefficient, point biserial correlation coefficient, and phi coefficient.[1,5] The basic purpose of the methods is to give a numerical value to the relationship between scores for the total MCQ test and the score for a single item. This numerical value is the index of the discriminatory effectiveness of the item. Although there are various similar ways of calculating the discrimination index, we used the simplified technique of selecting the upper and lower 27%, which have been demonstrated by Kelley[8] to be the most efficient fraction. The main limitation of the use of this method in estimating discrimination power is that it cannot be used for small sample size.

Discrimination indices are important in that poor discriminatory items are a valuable signpost towards ambiguous wording, grey areas of opinion and perhaps, even wrong keys. However, we must recognise that there may be other factors that need to be taken into account when using discrimination indices to categorise MCQs as "good" or "bad", especially when dealing with a multidisciplinary paper.[9] For example, students' performance in an MCQ item on Pharmacology may not accurately predict their performance in the MCQ items of another discipline, say Pathology, nor their overall performance in the total MCQ test scores of such a multidisciplinary test paper.

The wide scatter of item discrimination values for questions with a similar level of difficulty may reflect that some extent of guessing practices still occurred despite penalty marking. Test items with very poor discrimination indices should be reviewed by the respective disciplines. It serves as an effective feedback to the departments concerning their educational activities. When a test item appears to be very difficult (i.e., P is very small), it may be that the topic tested is inappropriate at this stage of students' training, or that it is not taught well or not taught at all in this particular academic session. Other possible reasons for poor performance on the items (i.e., D is very small) include ambiguity in the wording, areas of controversy, and perhaps, even that the wrong key was given. It is possible that a "good" student might not risk attempting a "difficult" MCQ item for fear of losing hard-earned marks on the other items of the same question. However, a "weak" student might take the risk to guess as he knows so little on the topic that he has nothing much to lose, and the least he can obtain for the whole question is zero marks. This could then result in a negative discrimination index.

It would be interesting to track the performance of students on the same MCQ items over time. Would this MCQ item have similar difficulty index and discriminatory index when tested in students from different cohorts who are at the same stage (e.g., Year II) of their medical training? This is difficult to evaluate in our study because the number of questions repeated in the subsequent tests is too small. Furthermore, 3 consecutive years may not be sufficient to make a reliable judgement on this.

It is interesting to note that despite the lack of written guidelines or the use of item analysis to help the lecturer in constructing the MCQ test items, a consistent level of test difficulty (and hence, standard) appears to be maintained from term to term and from year to year. The fact that any newly constructed MCQ has to go through several levels of vetting by peers, who are content experts as well as non-content experts, before its eventual use in an examination paper might have ensured this observed consistency. However, the wide scatter of discrimination among the MCQ items of similar level of difficulty can perhaps be substantially reduced by such an evaluation exercise on the item so that the quality of the standard of these MCQ tests can be further improved.

We hope the findings of this study will initiate a change in the way we select our future MCQ items, one of the several methods of assessment used in our undergraduate medical curriculum, and one on which our students appeared to consistently perform poorer compared to the other assessment methods, such as the SAQs and OSPEs. Based on the end-of-course self-evaluation by the students, time does not seem to be the major factor for the poorer performance in the MCQ test. The importance of evaluating assessment has been highlighted by Fowell and co-workers,[10] who noted that when devising suitable assessment systems, this step of the assessment cycle is often omitted. And yet most medical educators involved in curriculum planning and development recognise the interplay between assessment and learning, and that to a large extent assessment drives learning. Therefore, developing an appropriate assessment strategy is a key part of effective sustainable curriculum development.

Conclusions

There is a consistent spread of difficulty of MCQ (true/false-format) items in test papers across the different terms and years. MCQ items that demonstrate good discrimination tend to be in the moderately easy to moderately difficult range. On the other hand, items that are in the moderately difficult to very difficult range are more likely to show negative discrimination. The wide scatter of discrimination needs further investigation, and before we discard an MCQ for poor discrimination, we must first look into the factor(s) that may contribute to such poor discrimination.

Acknowledgements

The authors would like to thank the Dean of the Faculty of Medicine for his kind permission to review and analyse the results of the past MCQ tests, Miss Kamila Banu Rahim Khan for her assistance in data analysis, and my 3 colleagues (Dr Quek Kia Fatt, Professors Nor Azila Mohd Adnan and Geh Sooi Lin) for their valuable comments on the manuscript.

This study was funded by IRPA grant (07-02-03-1006) from MOSTI, Malaysia.

REFERENCES

1. Hubbard JP, Clemans WV. Multiple-choice Examinations in Medicine: A Guide for Examiner and Examinee. London: Lea & Fabiger, 1961.
2. De Champlain AF, Melnick D, Scoles P, Subhiyah R, Holtzman K, Swanson D, et al. Assessing medical students' clinical sciences knowledge in France: a collaboration between the NBME and a consortium of French medical schools. Acad Med 2003;78:509-17.
3. Norman G. Evaluation methods: A resource handbook. In: Shannon S, Norman, G, editors. Chapter 4.1. Multiple choice question. The Program for Educational Development, McMaster University. Hamilton, Canada: McMaster University, 1995:47-54.
4. Peitzman SJ, Nieman LZ, Gracely EJ. Comparison of "fact-recall" with "higher-order" questions in multiple-choice examinations as predictors of clinical performance of medical students. Acad Med 1990;65: S59-S60.
5. Backhoff E, Larrazolo N, Rosas, M. The level of difficulty and discrimination power of the Basic Knowledge and Skills Examination (EXHCOBA). Revista Electrónica de Investigación Educativa, 2000;2(1). Available at: http://redie.ens.uabc.mx/vol2no1/contents-backhoff.html. Accessed 27 November 2004.
6. Ross MM, McDonald B, McGuinness J. The palliative care quiz for nursing (PCQN): the development of an instrument to measure nurses' knowledge of palliative care. J Adv Nurs 1996;23:126-37.
7. Pellegrino J, Chudowsky N, Glaser R, editors. Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: National Academic Press, 2001.
8. Kelley TL. The selection of upper and lower groups for the validation of test items. J Educ Psychol 1939;30:17-24.
9. Hobsley M. Counting apples with oranges: a limitation of the discrimination index. Med Educ 1999;33:192-6.
10. Fowell SL, Southgate LJ, Bligh JG. Evaluating assessment: the missing link? Med Educ 1999;33:276-81.