Chula Med J Vol. 46 No. 11 November 2002

The repeated use of MCQ

Suwimol Sanpavat* Apichai Khongphatthanayothin*
Pornthep Lertsapcharoen* Pairoj Chotivitayatarakorn*

Sanpavat S, Khongphatthanayothin A, Lertsapcharoen P, Chotivitayatarakorn P. The repeated use of MCQ. Chula Med J 2002 Nov; 46(11): 907 - 15

Objectives : 1. *To assess the quality of MCQ in terms of levels of educational objective, the difficulty and the discrimination indices, and their classification based on the criteria set by WHO.*

2. *To explore effects on the scores made by 6th. year medical students, the difficulty and the discrimination indices to the subsequent uses of the same set of questions.*

Setting : *Department of Pediatrics, Faculty of Medicine, Chulalongkorn University.*

Design : *Retrospective , analytic study*

1. *The 5th medical year GPAX of the 6th year medical students.*
2. *The scores they earned from MCQ examination when they completed the rotation.*

Instrument : *Two hundred items of multiple choice questions.*

Method : *6th year medical students were divided into 2 groups according to the ranking of the test used. The 200 items were first used on student group 1, and repeated used for the second time on group 2. The format and quality of the test questions were assessed. Item analysis was calculated and classified according to WHO criteria. The difference of their GPAX, mean test scores, the difficulty and discrimination indices between groups were compared. Percentage, and unpaired t-test were used for statistic analysis.*

* Department of Pediatrics, Faculty of Medicine, Chulalongkorn University

Result : *The questions of recall accounted for 34.5 %, comprehension 21 % and problem- solving 44.5 %. All 200 items were one best answer type, with 62.5% contained clinical vignette, 36.5 % with negative lead-in, of which 7.5 % had double negatives. Thirty- nine percent of the questions were in the acceptable range for their difficulty index, 20.5 % was difficult and 40.5 % easy. Fifty- eight percent had good to excellent discrimination index, the rest were poor. Twenty- nine percent had the quality of combined acceptable difficulty index and good to excellent discrimination.*

The reliability of the tests were 0.52 and 0.51 in group 1 and 2 respectively. Comparison of GPAX between groups showed no difference. However, the mean test scores of group 2 who took the previously tested items were significantly higher than group 1 who took the original test ($p < 0.01$), The same result applied to the difficulty index. The repeated used items in group 2 obtained significant higher means than the original one in group 1 ($p < 0.001$). There was no difference of the discrimination index between groups.

Conclusion : *Two-thirds of this set of MCQ was to test comprehension and problem- solving. Twenty-nine percent of the test had good quality and could be used repeatedly. Student memorizing the used questions reduced the complex quality questions to be a recall, caused higher difficulty index of the test, and earned higher scores. Replacement of some obvious keywords and correct options without alteration of the question objectives should be done if they are repeatedly used.*

Keywords : *MCQ, Multiple choice question, Test , Scores, Item analysis, Item-bank.*

Reprint request : Sanpavat S, Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok 1030, Thailand.

Received for publication. September 16, 2002.

สุวิมล สรรพวัฒน์, อภิชัย คงพัฒนะโยธิน, พรเทพ เลิศทรัพย์เจริญ, ไพโรจน์ โชติวิทยธารากร. การใช้ข้อสอบ MCQ ซ้ำ ๆ กัน. จุฬาลงกรณ์เวชสาร 2545 พ.ย; 46(11): 907 – 15

วัตถุประสงค์ : *เพื่อศึกษา*

1. *คุณภาพของข้อสอบ MCQ ตามระดับวัตถุประสงค์การศึกษา การวิเคราะห์ข้อสอบรายข้อ และจำแนกตาม criteria ของ WHO ว่าด้วยการจำแนกข้อสอบ*
2. *ผลกระทบของการใช้ข้อสอบซ้ำ ๆ กัน ที่มีต่อคะแนนสอบของนิสิต ค่าความยากง่าย และอำนาจการจำแนกของข้อสอบ*

สถานที่ทำการศึกษา : *ภาควิชากุมารเวชศาสตร์ คณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย*

รูปแบบการวิจัย : *การศึกษาย้อนหลังเชิงวิเคราะห์*

กลุ่มตัวอย่าง : *- คะแนน GPAX ในระดับการศึกษาชั้นปีที่ 5 ของนิสิตแพทย์ปีที่ 6 ที่หมุนเวียนเข้าเรียนในภาควิชา และคะแนนสอบภายหลังการเรียนวิชากุมารเวชศาสตร์นาน 6 สัปดาห์*

เครื่องมือ : *- ข้อสอบชนิด MCQs (multiple choice questions) จำนวน 200 ข้อ*

วิธีการ : *- จัดกลุ่มนิสิตแพทย์เป็น 2 กลุ่มตามครั้งที่ของการใช้ข้อสอบซ้ำ ๆ กัน กลุ่มหนึ่งคือ นิสิตกลุ่มที่ใช้ข้อสอบครั้งแรกจำนวน 200 ข้อ กลุ่มสองใช้ข้อสอบชุดเดียวกันเป็นครั้งที่สองจำนวน 200 ข้อ*

- สำรวจแบบฟอร์ม และคุณภาพของข้อสอบโดยแยกเป็นข้อสอบระดับความจำ ความเข้าใจและการแก้ปัญหา

- วิเคราะห์ข้อสอบรายข้อ (item analysis) หลังการใช้สอบแต่ละครั้ง หาค่าความยากง่าย (p - difficulty index) และค่าอำนาจการจำแนก (r – discrimination index) จัดกลุ่มข้อสอบตาม criteria ของ WHO

- จัดกลุ่มคะแนน GPAX และคะแนนสอบ MCQ ตามกลุ่ม 1, 2

- วิเคราะห์ความแตกต่างของ GPAX คะแนนสอบ ค่าเฉลี่ยความยากง่าย และอำนาจการจำแนกระหว่างการใช้ครั้งที่ 1 และ 2 โดยใช้ค่าร้อยละ และ unpaired t-test

ผลการศึกษา : *คุณภาพของข้อสอบ 200 ข้อ เป็นการวัดระดับความจำ 34.5 %, ระดับความเข้าใจ 21% และ ระดับแก้ปัญหา 44.5 % ข้อสอบทั้งชุดเป็นชนิด one best answer. 62.5 % มีข้อมูลทางคลินิกในส่วนของคำถาม 36.5 % เป็นประโยคปฏิเสธ (negative lead-in) และ 7.5 % ของประโยคปฏิเสธเหล่านี้มีข้อความปฏิเสธซ้อนปฏิเสธ*

การวิเคราะห์ข้อสอบรายข้อ 39 % ของข้อสอบ มีระดับความยากง่ายที่อยู่ในเกณฑ์ยอมรับได้ตาม criteria ของ WHO 20.5 % อยู่ในเกณฑ์ยาก และ 40.5 % อยู่ในเกณฑ์ง่าย 58 % มีอำนาจในการจำแนกในเกณฑ์ดีถึงดีเยี่ยม (อำนาจการจำแนก $\geq$ 0.25) ข้อสอบที่เหลือมีอำนาจจำแนกต่ำ คือ $<$ 0.15 ข้อสอบที่มีความยากง่ายที่ยอมรับได้ร่วมกับมีอำนาจในการจำแนกมี 58 ข้อ คิดเป็น 29 % ความเชื่อมั่นของแบบทดสอบในการใช้ครั้งแรกเท่ากับ 0.52 และครั้งที่สองเท่ากับ 0.51 การเปรียบเทียบ GPAX ของนิสิตทั้งสองกลุ่มไม่มีความแตกต่างกัน แต่พบความแตกต่างกันของคะแนนเฉลี่ยที่ได้จากการใช้ข้อสอบชุดเดียวกัน โดยกลุ่มที่สอบซ้ำมีคะแนนเฉลี่ยสูงกว่ากลุ่มที่สอบข้อสอบเป็นครั้งแรกอย่างมีนัยสำคัญทางสถิติ ($p < 0.01$)

การวิเคราะห์ข้อสอบรายข้อ พบความแตกต่างกันของค่าความยากง่าย (difficulty index) เช่นเดียวกับคะแนนเฉลี่ย โดยข้อสอบที่ใช้ซ้ำมี difficulty index สูงกว่าข้อสอบที่ใช้ครั้งแรก ($p < 0.001$) แต่ไม่มีความแตกต่างกันในการเปรียบเทียบความสามารถในการจำแนก (discrimination index)

สรุป : *สองในสามของข้อสอบ MCQ ชุดที่นำมาศึกษาเป็นข้อสอบที่วัดความสามารถระดับความเข้าใจและแก้ปัญหา 29% มีระดับความยากง่ายและอำนาจในการจำแนกอยู่ในเกณฑ์ดี และสามารถเก็บไว้ในคลังข้อสอบเพื่อใช้สอบซ้ำได้ แต่การใช้ข้อสอบซ้ำ ๆ มีปัญหาเนื่องจากนิสิตจำข้อสอบได้ และทำให้การวัดความสามารถในระดับความเข้าใจหรือการแก้ปัญหากลายเป็นเพียงระดับความจำ การใช้ข้อสอบซ้ำควรมีการดัดแปลงแก้ไขถ้อยคำเฉพาะในส่วนของข้อมูลหรือเปลี่ยนแปลงตัวเลือกที่ถูกโดยไม่ทำให้มีการเปลี่ยนแปลงวัตถุประสงค์*

Multiple-choice question (MCQ), the objective-written test, is recognized as the most applicable and one among the popular tests in assessing students' knowledge.[1,2] Although there is a disadvantage of choices being provided for selection, which is not realistic in natural situations and therefore, subject to guessing; the advantages of sampling a broad range of topics and easily computerized - scored have made it frequently used both at the level of the department and the certified examinations. Construction of a good MCQ is crucial for the value of evaluation, but the task is not easy and is time consuming. Test banking and putting the good questions for repeated use, are advocated to increase the efficiency in using MCQ and improvement of their quality.[2]

With the periodic year-round examinations of medical students and postgraduate trainees at the Department of Pediatrics, the task of providing sets of well-written MCQs is a burden for the instructors. A particular set of MCQs was then developed for repeated use in the 6th year medical students' examination.

The objectives of this study are namely: 1) to assess the quality of the test in term of the educational levels, the difficulty and discrimination indices, and classify them according to the criteria set by WHO[3] (Table 1). 2) to explore the effect on the scores earned by 6th year medical students, the difficulty and the discrimination indices of the test, for the subsequent uses of the same MCQs. Because all the students in this study were in the same academic year, were exposed to the same pediatric curriculum and the same set of instructors, therefore it is assumed that this variable of the samples being rotated at different time period has little effect on their performance on the test.

Table 1. The criteria of WHO in classifying the test questions.[3]

difficulty index (p)	0.3 - 0.7	acceptable
discrimination index(r)	≥ 0.35	excellent
	0.25 - 0.34	good
	0.15 - 0.25	marginal (revise)
	< 0.15	poor (discard)

Method

At the end of a 6-week rotation to the Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, the 6th year medical students' knowledge was evaluated using MCQs on the topics which listed in the Pediatric Course Syllabus[4] to ensure the validity of its content. Two hundred items of MCQs were constructed. They were first used on students group 1, and reused for the second time on group 2 in the same academic year. The mean students' scores of each group were calculated. The GPAX (mean grade point average) of the 5th year of each student was collected and grouped. Item analysis of each question was calculated, using CTIA grading[5] after each examination.

The data were analyzed using the examinees' scores (as mean, standard deviation, GPAX of the 5th medical-year) and the test item [difficulty index (p) and discrimination index (r)] for the units of analysis. Descriptive results were calculated as percentage; and unpaired t-test was computed to determine the relationship between the groups.

Result

In assessing the quality of the test, it was discovered that 69 of 200 items (34.5 %) were in the educational level of recall, 42(21 %) comprehension and 89 (44.5 %) in problem - solving. All items were the one-best answer type. One hundred and twenty five items (62.5 %) contained data or clinical vignettes, 73 items (36.5 %) had negative lead-ins, of which 5 (7.5 %) also had double negatives in the options.

Item analysis revealed that 78 questions (39 %) had difficulty index (p) between 0.3-0.7; p <0.3 in 41 (20.5 %) and p > 0.7 in 81 (40.5 %). Fifty nine items (29.5 %) had excellent discrimination index ($r \geq 0.35$), 57 items (28.5 %) had r 0.25 - 0.34, and 84 questions (42 %) had r < 0.15. There was no test item that had r between 0.24 and 0.15 (Table 2). When taking both indices into consideration, 58 (29 %) had p 0.3-0.7 and $r \geq 0.25$; 47 (23.5 %) had p outside the acceptable range, but r was still ≥ 0.25.

The two groups of examinees were classified based on the repeated use of the test items. The number of the examinees were 72 and 69 in group 1 and 2 respectively. The reliability of the test were 0.52 in group 1 and 0.51 in group 2. Comparison of GPAX of the 5th medical-year between the two groups showed no statistically significant difference. However, a significant difference was observed on the mean test scores when the previously tested items were repeatedly used ($p < 0.01$), (Table 3).

Comparison of the mean difficulty index between these 2 groups also showed significant difference ($p<0.001$), but there was no difference in the discrimination index .

Table 2. Item analysis of 200 MCQs based on WHO criteria.

		no.	%
Difficulty index (p)			
> 0.7	easy	81	40.5
0.3 - 0.7	acceptable	78	39
< 0.3	difficult	41	20.5
Discrimination index (r)			
> 0.35	excellent	59	29.5
0.25 - 0.34	good	57	28.5
0.15 - 0.24	marginal (revise)	0	0
< 0.15	poor (discard)	84	42

Table 3. Comparison of GPAX, mean test scores, difficulty and discrimination indices between 2 groups of examinee when the MCQs were repeatedly used.

Criteria	Group 1	Group 2	P
No. of examinees	72	69	
No. of test items	200	200	
Reliability	0.52	0.51	
GPAX	3.18	3.16	0.6
Test scores			
Mean (SD)	28.54 (4.75)	30.80 (4.44)	< 0.01*
Difficulty index (p)			
Mean (SD)	0.57 (0.07)	0.62 (0.08)	< 0.001*
Discrimination index (r)			
Mean (SD)	0.22 (0.31)	0.22 (0.29)	0.75

Discussion

MCQ that can measure complex ability, such as comprehension and application of knowledge to the patient care is preferable to measuring the recall of isolated facts. Two- thirds of our items (65 %) contained questions involving comprehension and application of knowledge. The amount of these complex questions should be raised to verify students' clinical competency and sounded judgment. How to construct a good quality MCQ has been suggested.[6,7,8] Vignette item with no flaw and one-best answer format are rated the highest in the rating scheme.[9] Our set of MCQs has met most of these qualities. The questions are homogeneous, only comprised of one-best-answer type and are grouped into subject areas. Sixty-two percent have data or patient vignettes in the stems, which enable the examinees to recognize the nature of the desired responses without figuring out from the options. However, the weakness of the test is in 73 questions (36%) that are expressed in the negative terms, such as: not, less likely, the least, and except. This may introduce an ambiguity, unwanted confusion, and misleading, since students are likely to seek the true rather than the false statements. Only 2 items have either the distracters like "all of the above" or "none of the above" which should be avoided. Factors that undeniably effect the quality of the test are there is no review of questions before they are put into use and the pattern and format of clinical vignette that stress on no negative phrases are not always followed.

Item analysis provides information to identify deficiency in questions.[5] Our study reveals 39% as acceptable questions, for their difficulty index of 0.3-0.7 is likely to be reliable regarding its internal consistency or homogenicity.[6] Forty percent of the questions with the difficulty index > 0.7 are considered easy and 20% with the difficulty index < 0.3 are

considered difficult. More than half of the questions (58 %) have good to excellent discriminating power ($r>0.25$). Twenty- nine percent have acceptable difficulty combined with good discrimination and are suitable for banking. Considering the essential knowledge with reference to the educational objectives, some very easy or difficult questions are also important and should not be discarded. These may partly fall into the 23.5 % that has discriminating power but not yet achieved the desired level of difficulty. Revision is needed to improve their quality.

To be fair to the students across the academic year, the difficulty of the exam should be kept uniform and consistent. The repeated use of the well-written questions supposedly serves this purpose, but the question of security has tarnished its determination. It is also every student's desire to get as good marks as possible. All too often, groups of students will systematically memorize the questions, and emphasize only the keywords of the stems and their correct responses, and hand over them to their peers. Thus the intended comprehension and decision- making objectives have been reduced to merely a recall. Our results provide the same evidence. The students' GPAX of a previous-year were not different, but when groups 2 was tested with the previously used questions, the mean scores were significantly higher when compared to group 1 which used the original items. This is also true when the mean difficulty index was compared. The repeated use of the test was easier and more students in groups 2 answered correctly than group 1 , resulting in higher difficulty index ($p < 0.001$). The easy way for the student to memorize as much items as possible is to stress only on the keywords in the stems and their correct choices. To confront this problem, replacement has to be made, but it has to be done closely relevant to the original objective. Since the distracters or the incorrect options are not the subjects of interest to the students who memorize the clues, the alteration has to be made in the stem and the correct response, which is not easy, and by no mean it is possible to be always correlated with the preset objective.

Conclusion

All questions are the one-best answer type, with two-third to test the ability of comprehension and problem-solving. Twenty-nine percent of the items met the WHO criteria in classifying as acceptable difficulty index and in the range of excellent and good discrimination, thus are suitable for item banking. Efficient use of the well-written MCQs by repeated apply to different groups of students can be done but with some replacement that is relevant to the preset objective.

Acknowledgement

We would like to thank Lertmaharitt S, Pholwan N. and Laisnitsarekul B. for their statistic advice and information.

References

1. Gronlund NE, Lin RL. Constructing objective test items : multiple-choice forms. In: Miller R,ed. Measurement and Evaluation in Teaching. 6th ed. New York : Macmillan. 1990: 166-191
2. Bandaranayake R. How to organize multiple choice question banks. In : Cox KR, Ewan CE, eds. The Medical Teacher. 2nd ed. London:

Churchill, Livingstone, 1988: 1966 - 9

3. Guilbert JJ. Critical evaluation of a question. In: Educational Handbook for Health Personnel. 6th ed. WHO offset Publication No. 35.Geneva: WHO, 1998: 4.80 - 4.81
4. Chulalongkorn University. Pediatric Course Syllabus for the 6th year Medical Students. Department of Pediatrics, Faculty of Medicine, Chulalongkorn University. 1999.
5. Sukamolson S. CTIA Grading, Classical Test Item Analysis and Grading Manual (V6.30). Bangkok :Chulalongkorn University Language Institute, 1992: 36
6. Weber ML. Evaluation of learning : an important task ? Annals 2002 (April).The Royal College of Physician and Surgeon of Canada, 2002.
7. Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Science. 2nd ed. Philadelphia: National Board of Medical Examiner, 1998.
8. Gay LR. Constructing test items. In : Gay LR, ed. Educational Evaluation and Measurement. 2nd ed. New York: Macmillan, 1985: 217 - 39
9. Jozefowiez RF, Bruce KM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med 2002 Feb; 77(2) : 156 - 61
10. Bandaranayake R. and Cox K. Writing multiple choice questions. In: Cox KR, Ewan CE, eds. The Medical Teacher. 2nd ed. London: Churchill Livingstone, 1988: 152 - 6