Chula Med J Vol. 47 No. 4 April 2003

Comparing reliabilities before and after reducing recall items in Comprehensive Examination MCQ tests*

Boonnart Laisnitsarekul**

Vitool Lohsoonthorn*** Oranuch Kyokong****

Laisnitsarekul B. Lohsoonthorn V, Kyokong O. Comparing reliabilities before and after reducing recall items in Comprehensive Examination MCQ tests. Chula Med J 2003 Apr; 47(4): 231 - 40

Objective : *To compare the reliabilities of the Comprehensive Examination MCQ tests of the Academic Years 1994, 1997, 1999, 2000 and 2001 before and after reducing recall items.*

Design : *Descriptive study.*

Methods : *Comprehensive Examination MCQ tests of the Academic Years 1994-2001 were composed of the following sections: recall, interpretation and problem-solving categories. Their reliabilities were calculated before and after reducing the recall category. The reliability of MCQ tests was tested for statistics significance by Nonparametric statistics, randomization test for match pairs.*

Results : *The reliabilities of Comprehensive Examination MCQ tests of the Academic Years 1994, 1997, 1999, 2000 and 2001 were 0.84, 0.87, 0.86, 0.84 and 0.84, respectively. Being classified into three categories, recall, interpretation and problem- solving, MCQ tests were composed of recall items 5 %- 25 %, interpretation items 19 % - 32 % and problem-solving items 51 %-72 %. The differences between the reliabilities of MCQ tests, before and after reducing recall items, were not statistically significant.*

---

* In the name of the 2002 Comprehensive Examination Committee

** Medical Education Unit, Faculty of Medicine, Chulalongkorn University

*** Department of Preventive and Social Medicine, Faculty of Medicine, Chulalongkorn University

**** Department of Anesthesiology, Faculty of Medicine, Chulalongkorn University

Conclusion : *Comprehensive Examination MCQ tests contain a number of contents since 1996, being constructed based on the Table of Specification. This study found that the reliabilities of MCQ tests of the Academic Years 1994, 1997, 1999, 2000 and 2001, before and after reducing the recall items, were not of statistical difference. The Comprehensive Examination Committee could reduce the number of multiple-choice items by constructing the interpretation and problem-solving items based on the table of specification. This research has shown that the number of multiple-choice items could be less than 300 and yet their reliability can remain high. The committee can also save time, materials and energy in constructing and managing MCQ tests.*

Keywords : *Reliability, Comprehensive Examination, MCQ, Multiple-choice, Reducing, Recall item.*

Reprint request: : Laisnitsarekul B. Medical Education Unit, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand.

Received for publication. January 12, 2003.

บุญนาท ลายสนิทเสรีกุล, วิฑูรย์ โล่ห์สุนทร, อรนุช เกี่ยวข้อง. การเปรียบเทียบความเที่ยงก่อนและหลังการคัดแยกข้อสอบประเภทความจำ ในข้อสอบปรนัย วิชาเวชศาสตร์ทั่วไป. จุฬาลงกรณ์เวชสาร 2546 เม.ย; 47(4): 231 – 40

วัตถุประสงค์ : *เพื่อเปรียบเทียบความเที่ยงของข้อสอบปรนัย วิชาเวชศาสตร์ทั่วไป ปีการศึกษา 2537, 2540, 2542, 2543 และ 2544 ก่อนและหลังการคัดแยกข้อสอบประเภทความจำ*

รูปแบบการวิจัย : *การศึกษาเชิงพรรณนา*

วิธีการศึกษา : *ข้อสอบปรนัย วิชาเวชศาสตร์ทั่วไป ปีการศึกษา 2537, 2540, 2542, 2543 และ 2544 ถูกนำมาจำแนกรายข้อว่าเป็นข้อสอบประเภทความจำ ความเข้าใจและการแก้ปัญหา ข้อสอบปรนัยทุกฉบับนำมาคำนวณหาค่าความเที่ยงทั้งก่อนและหลังการคัดแยกข้อสอบประเภทความจำ และนำค่าความเที่ยงทั้งสองชุด มาเปรียบเทียบหาความแตกต่างทางสถิติ ด้วยสูตรสถิตินอนพาราเมตริก, แรนด้อมไมเซชั่น เทส ฟอร์ แม็ช แพร์ (Randomization test for match pairs).*

ผลการศึกษา : *ความเที่ยงของข้อสอบปรนัย วิชาเวชศาสตร์ทั่วไป ปีการศึกษา 2537, 2540, 2542, 2543, และ 2544 มีค่าเท่ากับ 0.84, 0.87, 0.86, 0.84, และ 0.84 ตามลำดับ เมื่อนำมาจำแนกเป็นรายข้อ มีข้อสอบประเภทความจำร้อยละ 5 - 25 ข้อสอบประเภทความเข้าใจร้อยละ 19 - 32 และข้อสอบประเภทการแก้ปัญหาร้อยละ 51-72 เมื่อเปรียบเทียบความแตกต่างของความเที่ยงข้อสอบปรนัย ก่อนและหลังการคัดแยกข้อสอบประเภทความจำ ไม่พบความแตกต่างอย่างมีนัยสำคัญทางสถิติ*

สรุป : *ข้อสอบปรนัย วิชาเวชศาสตร์ทั่วไป มีความตรงเชิงเนื้อหา มาตั้งแต่ปี พ.ศ. 2539 เนื่องจากข้อสอบถูกสร้างตามตารางการวิเคราะห์เนื้อหา (Table of Specification) การศึกษาครั้งนี้พบว่า ความเที่ยงของข้อสอบในปีการศึกษา 2537, 2540, 2542, 2543 และ 2544 ก่อนและหลังการคัดแยกข้อสอบประเภทความจำ ไม่มีความแตกต่างอย่างมีนัยสำคัญทางสถิติ คณะกรรมการบริหารการสอบ วิชาเวชศาสตร์ทั่วไป สามารถลดจำนวนข้อสอบปรนัยโดยการสร้างข้อสอบประเภทความเข้าใจและการแก้ปัญหาตามตารางการวิเคราะห์เนื้อหา งานวิจัยนี้แสดงให้เห็นว่า จำนวนข้อสอบปรนัยสามารถมีจำนวนน้อยกว่า 300 ข้อ โดยที่ความเที่ยงยังมีค่าสูง คณะกรรมการบริหารการสอบวิชาเวชศาสตร์ทั่วไป สามารถประหยัดเวลา วัสดุ และกำลังงานในการสร้างและบริหารการสอบข้อสอบปรนัยได้อีกด้วย*

Multiple-choice questions are the most flexible objective items type. They can be used to appraise the achievement of any educational objective that can be measured by a paper-and-pencil test, except those related to skills in written expression, originality and the ability to organize a response. An ingenious and talented item writer can construct multiple-choice items that require not only recollection of knowledge but also the use of skills for comprehension, interpretation, application, analysis, or synthesis to arrive at the correct answer.[1]

1985 was the first year when sixth year medical students, Faculty of Medicine, Chulalongkorn University of Academic Year 1979 were required to pass the Comprehensive Examination before receiving their medical licenses. The measuring instrument was MCQ which be constructed under the Table of Specification and Thai Medical Council's Standard Criteria for content validity since 1996.[2] Each item must be referable to the established criteria. (A sample is shown in Figure 1.) Based on the curriculum development MCQ test trends to construct on interpretation and problem-solving.[3] Medical students must complete the university's criteria [4] and pass the Comprehensive Examination Committee's judgment. [5]

Multiple-choice Form

1. **Department** Ophthalmology
2. **Division or Unit** -
3. **Issuer** -
4. **Condition or Disease** Ophthalmic drug
5. **Objectives** Identify the ophthalmic drug which cause of open-angle glaucoma if prolonged use.
6. **Standard** Table of Specification No. 2.1.1
7. **Level** Interpretation
8. **Type** One-best
9. **Difficulty Factor (D.F.)** 0.60
10. **Question** Which of the following ophthalmic drugs, in prolonged use can cause open-angle glaucoma ?
    - A. Topical anesthetics.
    - B. Topical corticosteroids.
    - C. Chloramphenicol eye drop.
    - D. Beta – adrenergic blocker.
    - E. Epinephrine eye drop.
11. **Answer** B
12. **Ref.** Vaughan D, Asbury T. General Ophthalmology. 10th ed. 1983
13. **p value** 0.65
14. **r value** 0.36

Figure 1. A sample of Comprehensive Examination MCQ which is based on the Table of Specification.

The Comprehensive Examination MCQ test of 1979 contained 500 items: 322 (64.40 %) were recall items; 59 (11.80 %) interpretation items; 119 (23.80 %) problem-solving items.[6] The Comprehensive Examination MCQ test of 1994 had 300 items. The students had to finish their test in two days; each day they had to answer 150 items in three hours[7]; this was similar to the Thai Medical Council's MCQ test on clinical sciences. It had 450 items divided into three parts; each part had 150 items to be answered in three hours.[8] Charvat, McGuire and Parsons [9] in the names of World Health Organization pointed out that MCQ's construction is time-consuming if arbitrary and ambiguous questions are to be avoided. Then the authors who are interested to find out the way to reduce the number of Comprehensive Examination multiple-choice items and kept the reliability in high level.

The purpose of this study is to compare the reliability of the Comprehensive Examination MCQ tests in Academic Years 1994, 1997, 1999, 2000 and 2001 before and after reducing the recall items.

Materials and methods

The research design in this study is descriptive. The study materials are the Comprehensive Examination MCQ tests of the Academic Years 1994, 1997, 1999, 2000 and 2001, Faculty of Medicine, Chulalongkorn University. The tests are classified into three categories, namely: recall, interpretation and problem-solving. They are also calculated for reliability before and after reducing the recall category. The reliability of MCQ tests is tested for statistics significance by nonparametric statistics, named randomization test for match pairs.

Results

1. The reliabilities of the Comprehensive Examination MCQ tests of the Academic Years 1979, 1994, 1997, 1999, 2000 and 2001 were 0.85, 0.84, 0.87, 0.86, 0.84 and 0.84, respectively. [Table 1]

2. The MCQ of the Comprehensive Examination, Academic Year 1979, were composed of recall, interpretation and problem-solving categories: 322 (64.40 %) were recall items; 59 (11.80 %) interpretation items; 119 (23.80 %) problem-solving items. In the Comprehensive Examination MCQ test of 1994, there were 37 (12.33 %) recall items, 97 (32.33 %) interpretation items and 166 (55.34 %) problem-solving items. In the Comprehensive Examination MCQ test of 1997, there were 75 (25.00 %) recall items, 72 (24.00 %) interpretation items and 153 (51.00 %) problem-solving items. In the Comprehensive Examination MCQ test of 1999 there were 68 (22.67%) recall items, 68 (22.67 %) interpretation items and 164 (54.66 %) problem-solving items. In the Comprehensive Examination MCQ test of 2000, there were 53 (17.67 %) recall items, 59 (19.67 %)

Table 1. Reliabilities of the Comprehensive Examination MCQ tests of the Academic Years 1979,1994, 1997,1999, 2000 and 2001.

Academic Year	Number of items	Reliability
1979	500	0.85
1994	300	0.84
1997	300	0.87
1999	300	0.86
2000	300	0.84
2001	299	0.84

Table 2. Table of Specification for Comprehensive Examination MCQ test.

Academic	Recall		Interpretation		Problem-solving		Total	
Year	f	%	f	%	f	%	f	%
1979	322	64.40	59	11.80	119	23.80	500	100.00
1994	37	12.33	97	32.33	166	55.34	300	100.00
1997	75	25.00	72	24.00	153	51.00	300	100.00
1999	68	22.67	68	22.67	164	54.66	300	100.00
2000	53	17.67	59	19.67	188	62.66	300	100.00
2001	17	5.69	68	22.74	214	71.57	299	100.00

Table 3. Comparative of reliabilities of the Academic Years 1994-2001 Comprehensive Examination MCQ tests before and after excluding the recall items by randomization test for match pairs.

Academic	Before		After		Significant
Year	Item	Reliability	Item	Reliability	
1994	300	0.84	263	0.83	NS
1997	300	0.87	225	0.82	NS
1999	300	0.86	232	0.82	NS
2000	300	0.84	247	0.83	NS
2001	299	0.84	282	0.83	NS

interpretation items and 188 (62.66 %) problem-solving items. In the Comprehensive Examination MCQ test of 2001, there were 17 (5.69 %) recall items, 68 (22.74 %) interpretation items and 214 (71.51 %) problem-solving items. [Table 2]

3. Comparing the reliabilities of the Academic Years 1994 -2001 Comprehensive Examination MCQ tests, before and after excluding the recall items by the randomization test for match pairs, they were not of statistical difference. [Table 3]

Discussion

When constructing or selecting assessments, the most essential characteristics are validity, reliability, and usability.[10] The Table of Specification for the Comprehensive Examination MCQ test covered subjects of clinical sciences and emphasized on problem-solving ability. Since 1996, the Comprehensive Examination MCQ tests contain varieties of contents, being based on the Table of Specification.The Chulalongkorn Medical Curriculum has been developed to provide more efficient doctors for the community hospital who have abilities to learn by themselves and should have their own responsibility and self-confidence under the good humanity with moral and medical ethics. The newly developed curriculum (1999) has its own philosophy that emphasizes the way of inquiry and problem-based learning (PBL).[11]

The Comprehensive Examination MCQ test trends to have more interpretation and problem-solving categories and the medical students cannot do the examination on time. The Comprehensive Examination Committee has tried to reduce the number of multiple-choice, increase the proportion of problem-solving category and retain the level of reliability since academic year 1979. Stodala and Stordahl[12] recommended that Table of Specification or two-way grid is a useful tool that could help a test writer prepare items consistent with his instructional objectives. It classifies each test item in accordance with two basic dimensions of each item, usually course content covered and student behavior required. Example of "content" and "behavior" are given in Table 4.

Based on the purpose for reducing the number of multiple-choice items, this study found that by cutting the recall items the reliability of the Comprehensive Examination MCQ tests did not have statistical difference. However, the validity of the test would be questionable. A way to solve this problem is by constructing items that have more combined complex objectives such as recall, interpretation and problem-solving. Bloom, Chausow, Dressel and Mayhew [13] suggest that there are some evidences that the test question intended to evaluate the objectives which fall in the higher (and more complex) parts of the cognitive domain are more difficult than the test questions intended to evaluate the less complex objectives. Bloom, Engelhart, Hill, Furst and Krathwohl[14]

Table 4. Table of Specification for the 2001 Comprehensive Examination MCQ test.

Content	Behaviors			Total
	Recall	Interpretation	Problem-solving	
Medicine	-	7	51	58
Pediatrics	6	13	29	48
Preventive and Social Medicine	2	12	6	20
Psychiatry	2	3	5	10
Ophthalmology	-	1	5	6
Forensic Medicine	-	2	3	5
Rehabilitation Medicine	-	-	3	3
Surgery	3	10	54	67
Obstetrics and Gynecology	1	8	35	44
Orthopedics	1	6	8	15
Radiology	1	5	3	9
Anesthesiology	1	-	7	8
Oto-laryngology	-	1	5	6
Total	**17**	**68**	**214**	**299**
	(5.69%)	**(22.74%)**	**(71.57%)**	**(100.00)**

explain that the behavior in interpretation is when given a communication the student can identify and comprehend the major ideas which are included in it as well as understand their interrelationships. The whole cognitive domain of the taxonomy is arranged in a hierarchy, i.e., each classification demands skills and abilities that are lower in the classification order. A demonstration of "comprehension" shows that students can use the abstraction when its use is specified. A demonstration of "application" shows that they will use it correctly, given an appropriate situation in which no mode of solution is specified. A multiple-choice question can be constructed on an item that covers both the recall and interpretation objectives. A sample is shown in Figure 2.

Multiple-choice Form

1. **Department** Preventive and Social Medicine
2. **Division or Unit** -
3. **Issuer** -
4. **Condition or Disease** Epidemiology
5. **Objectives** 5.1 Identify the type of study design.
   5.2 Tell the formula of relative risk.
   5.3 Interprete the result of relative risk.
6. **Standard** Thai Medical Council's Standard No. 1.1.1
7. **Level** Interpretation
8. **Type** One-best
9. **Difficulty Factor (D.F.)** 0.40
10. **Question**

In the study of the relationship between factor X and myocardial infarction (MI), the result of the study is shown in the table below:

At the beginnings of the study	Outcome		Total
	Developed MI	Did not develop MI	
Persons with factor X	300	700	1000
Person without factor X	100	1900	2000
Total	400	2600	3000

Which of the following is CORRECT ?

A. Factor X protects againt development of myocardial infarction.
B. There is no association between factor X and myocardial infarction.
C. There is a positive association between factor X and myocardial infarction.
D. There is either no association or a negative association between factor X and myocardial infarction.
E. There is either no association or a weak positive association between factor X and myocardial infarction.

11. **Answer** C
12. **Reference**

Figure 2. A sample of Comprehensive Examination MCQ which covered many levels of cognitive objectives.

Multiple-choice questions are one of the most widely applicable test items for measuring achievement. It can effectively measure various types of knowledge and complex learning outcomes. Another commonly cited advantage of a multiple-choice question is its greater reliability per item. As the number of alternatives increases from two to four or five, the opportunity for marking the correct answer is reduced; hence the reliability is correspondingly increased.[15] Hubbard and Clemans,[16] Schumacher[17], Cox and Ewan[18] have suggested that a good test should have a reliability of 0.70 or over. Regarding validity assurance, the Table of Specification has been introduced as basis for constructing items and tasks. This is merely a matter of checking to see whether the item or task is still relevant to the same cell in the table. As a rough guideline, [19] an average high school student should be able to answer one multiple-choice item per minute during the test. Interpretive test items normally take longer time; the exact amount of time spent depends on the length and complexity of the introductory materials. Since 1979, the Thai Medical Council's Standard Criteria and medical curriculum has based on interpretation and problem-solving ability; hence, the Comprehensive Examination Committee has tried to reduce the number of multiple-choice items and increased the proportion of the interpretation and problem-solving items. If the test has less than 300 items assigned for an examination of 3 hours, the committee can save time, materials and energy in its constructing, preparing, administering as well as appraising the results. Medical students will also have more time to do the interpretive and problem-solving multiple-choice items. However, the reduction of recall items is a way to serve the purpose.

References

1. Thorndike RM. Measurement and Evaluation in Psychology and Education. 6th ed. New Jersey: Prentice-Hall, 1997: 453
2. Tantayaporn K. Introduction. in : Tantayaporn K. Table of Specification for Comprehensive Examination and Comprehensive Examinations Handbook. Bangkok : Division of Academic Affairs, Faculty of Medicine, Chulalongkorn University, 1996; (a)
3. Chulalongkorn University, Faculty of Medicine. MD. Curriculum (1999 reformed curriculum). Bangkok : Division of Academic Affairs, Faculty of Medicine, Chulalongkorn University 1999:3
4. Chulalongkorn University. The Chulalongkorn University's Regulation and Announcement Based on Bachelor Degree Education. Bangkok: Chulalongkorn University Press, 2002; 1 - 15
5. Chulalongkorn University, Faculty of Medicine. Annoucement for Setting the 2002 Comprehensive Examination Committee. July 26, 2002
6. Phulklongtan M, Jaroongdaechakul M, Limpapayom K. An analysis of comprehensive examination items of academic year 1979. Chula Med J 1981 Sep; 25(9): 1035 - 40
7. Laisnitsarekul B, Tantayaporn K, Sriratanaban J. The 1993 Comprehensive examination in medicine : scores of multiple choice question tests. Chula Med J 1994 May; 38 (5): 279 - 91
8. Ministry of Public Health. Regulation for Medical License Examination, Thai Medical Council 1995. Bangkok : Thai Medical Council Permanent Secretary Office, 1995: 5 - 7

9. Charvat J, McGuire C, Parsons V. A Review of the Nature and Uses of Examinations in Medical Education. Geneva : World Health Organizaiton, 1968: 23

10. Linn RL, Gronlund NE. Measurement and Assessment in Teaching. 8th ed. New Jersey: Prentice-Hall, 2000: 73

11. Kamol-ratanakul P. Faculty of Medicine, Chulalongkorn University, Thailand. In: Stijnen M, Vluggen P, eds. The Network Newsletter. 2001 Dec; 36: 5 - 6

12. Stodola Q, Stordahl K. Basic Educational Tests and Measurement. New Delhi : Thomson Press (India), 1972: 15 - 9

13. Krathwohl DR, Bloom BS, Masia BB. Taxonomy of educational objectives : the classification of educational goals. In: Handbook II : Affective Domain. New York : David McKay, 1964: 8 - 12

14. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of educational objectives : the classification of educational goal. In: Handbook I : Cognitive Domain. 17th ed. New York : David McKay, 1972: 89-120

15. Linn RL, Gronlund NE. Measurement and Assessment in Teaching. 8th ed. New Jersey: Prentice-Hall, 2000: 200 - 2

16. Hubbard JP, Clemans WV. Multiple Choice Examinations in Medicine : A Guide for Examiner and Examinee. Philadelphia: Lea & Febiger, 1961: 71

17. Schumacher CF. Scoring and analysis. In: Hubbard JP, ed. Measuring Medical Education. Philadelphia : Lea & Febiger, 1971: 60 - 1

18. Cox K, Ewan CE. The Medical Teacher. 2nd ed. London : Churchill Livingstone, 1988: 163

19. Linn RL, Gronlund NE. Measurement and Assessment in Teaching. 8th ed. New Jersey: Prentice-Hall, 2000: 352 - 3