Final Test Analysis Of Post Graduate Medical Residents

Arab M, MD[1]; Davaloo Sh, MD[2]; Nasrolahi Sh, MD[3]; Nadi E, MD[4]; Emdadi M, MD[5]

[1] Assicuate professor, OB & GyN Departement, Hamadan University of medical sciences
[2] Educational instructor, OB & CyN Department, Fatemieh hospirtal of Hamadan
[3] Assistan professor, OB & Cy N Department Hamadan University of medical sciences
[4] Assistant professor, Internal medicine department, Hamadan University of medical sciences
[5] Assistant professor, pediatrics department, Hamadan University of medical sciences

Received: January, 2005 ***Accepted: May, 2005***

Abstract

Background and purpose: *Multiple choice questions are the most frequent test for medical students. It is important to analysis the overall response to individual questions in the test. The aim of this study is to analyse questions of post graduate medical residency tests.*

Methods: *Final annual local (Hamadan medical school) and national tests given to three Residency groups including 17 Obstetrics and gynecology testees, 7 pediatrics and 12 internal medicine in 2004 were studied. In local tests residents answered to 148, 150 and 144 and in national tests to ISO MCQS. Questions were evaluated regarding cognitive domain level, Difficultly index and Discriminative index and finally to evaluate the optimal, proper, acceptable and "must omitted" questions.*

Results: *Questions of local Obstetrics and gynecology, pediatrics and internal medicine tests evaluated the "recall" level in 72% , 72% and 51% and in national tests 71%, 35% and 19%, respectively. Questions with Discriminative indices of 0.7 or more (proper) were 3 and 5% in Obstetrics and gynecology, 3.5% and 1% in pediatrics and 1% in local and national tests. Proper difficulty indices (30-70) were shown in 53% and 54% in Obstetrics and gynecology, 34% and 43% in pediatrics and 40% and 42% in internal medicine. Generally evaluating, "must omitted" questions in local and national tests were 76% in Obstetrics and gynecology, 81% and 79% in pediatrics and 91% and 85% in internal medicine. The most common causes making the questions to be considered "must omitted" in studied tests were negative, zero or less than 0.2 Discriminative indices.*

Conclusion: *Test analysis of final annual local (Hamadan medical school) and national tests of Obstetrics and gynecology , Pediatrics and internal medicine residency programs in 2004 revealed that most of the questions are planned in "recall" level, harbor improper Discriminative indices, Difficulty indices and generally evaluating are "must omitted".*

Key word : *Multiple choice questions/ Test Analysis, Difficulty index/ Discriminative index*

Journal of Medical Education **Summer** ***2005; 7(2);67-72***

Introduction

The most important method to evaluate academic achievement is giving a test.(1)
Common academic achievement tests, known as paper and pencil tests are mostly performed in cognitive domains. Multiple choice tests are the most common objective tests, which are the best with respect to the question homogeneity, but have low sensitivity in identifying blind guess and easy marking (1).
Faced with the challenge to develop models of assessment relevant to work of physicians, in a study in college of medicine, Michigan State University, they concluded that a variety of competency assessments currently are used. MCQS remain a core assessment method (2).

Corresponding author: *Maliheh Arab, M.D. Fatemieh Hospital, pasdaran (Kerman shah) Street, Hamadan, Iran*
Tel: 0912-1593277
E-mail: drmarab@yahoo.com

A survey on data sources for decisions on medical students pass to higher levels indicated that multiple choice questions (MCQS) remains the most frequently used information source. Several explanations are suggested for the predominance of MCQS in students' evaluation, including familiarity for faculty and students, ease of implementation and the resources required for adoption of other assessment strategies (3).

In order to improve the quality of tests, one by one analysis of question and determination of their strengths and weaknesses are used (1).

Calculation of discrimination index, difficulty index and the level of cognitive domain evaluated in the questions could be used as a measure of quality.

Difficulty index is defined as the percent of examinees who answer a question correctly. The higher the difficulty index (near to 100), the easier the question and the lower the index (near to zero), the more difficult the question. The appropriate difficulty index is 0.6. The Difficulty indices between 0.3-0.7 are used to differentiate examinees (1).

Discrimination index determines power of the question to discriminate strong and weak examinees. The higher the discrimination index, the stronger the question and vice versa. Questions with Zero index can not discriminate examinees at all. Negative index simulate the weak examinees better than strong ones. These questions are basically considered improper and should be discarded (1).

Proper questions require a moderate difficulty index and high discrimination index.

In an approach to improve test construction, a survey was done in the George Washington university school to test item difficulty related to item discrimination. The results revealed the increased difficulty above a certain threshold did not produce a commensurate rise in discrimination, and was associated with many high- scoring students choosing incorrect answers for such items (4).

It is important (although time consuming in a paper- based assessment) to analyse the overall response to individual questions in the test, although drawing definite conclusions can be difficult (5).

This survey is to analyse questions of post graduate medical residency tests, regarding cognitive domain levels, difficultly index and discrimination index and finally to evaluate the proper and improper questions.

Methods

Final annual university based (Hamadan medical school) and national tests given to three seperate residency groups including 17 obstetrics and Gynecology (Obstetrics and gynecology), 12 internal medicine and 7 pediatrics residents of Hamadan University of Medical Sciences in 2004 were studied.

In these tests cognitive domain was evaluated using written multiple choice questions.

All residents of the 3 groups sat for the university based tests in May 2004 and answered to 148, 150, 144 multiple choice questions of their relevant subjects, respectively, and 2 months later in June 2004 took part in a national exam with 150 questions for each group.

The questions were divided into two groups according to their level of cognitive domain including:

1- "Recall" questions, directly assess recall and recognition of the students.

2- "Application" questions are measuring other levels of cognitive domain namely comprehension, application, analysis, synthesis and evaluation. Answers of these questions are not directly found in text books.

Discrimination index of questions was calculated one by one according to the
following formula:

DI of 0.7 or more was regarded proper.

Difficulty index was calculated, as follows:

PI between 30%-70% was regarded as proper. Indices of more than 70% and less than 30% indicate easy and difficult questions, respectively.

Final operational descriptions used in the evaluation of questions are as follow:

Optimal questions were defined as those questions with difficulty indices of 30-70, discrimination indices more than 0.7 and "application" type.

Discrimination Index (DI) = correct choices of strong group- correct choices of weak group.
Number of examinees in each group (strong or weak)

Difficulty index (PI) = correct choices of strong group + correct choices of weak group
Total number of examinees in strong group and in weak group

Proper questions were defined as those questions with difficulty indices of
30-70, discrimination indices more than 0.7 regardless of their level in cognitive domain.
Acceptable questions met one of the following criteria:
1- Difficulty indices of 20-80, discrimination indices of more than 0.5 regardless of their level in cognitive domain.
2- Difficulty indices of 20-80, discrimination indices of more than 0.2 and application type
3- "must omitted" questions met one of the following criteria):
1-Discrimination indices < 0.2 .
2-Discrimination indices between 0.2-0.5 coexisting with Difficulty indices of more than 80 (very easy) or less than 20(very difficult).
3- Discrimination indices between 0.2-0.5 and Difficulty indices of 20-80 in "recall" level.
Questions of local and national tests of 3 groups were compared based on these criteria.

Results

Questions of both local and national Obstetrics and gynecology residency group tests mostly evaluated the "recall" level, 72% and 71% respectively.(P=0.85)
These figures were 72% and 35% in pediatric group and 51% and 19% in internal medicine, respectively. A significant difference was observed in these two groups. (P<0.001) (Table 1)
Discrimination indices of local and national test questions were similar in all 3 residency groups with DIs of 0.7 or more found in 3% and 5% of Obstetrics and gynecology group questions and negative or zero indices, found in 35.5% and 33% respectively.
Questions with DIs of 0.7 or more in pediatric group were 3.5% and 1% and in internal medicine group were 1% in the both tests.
Negative or zero DIs in local and national tests were 60% and 54% in pediatric group and 47% and 52% in internal medicine group.(table 2)
Proper difficulty indices (30-70) of local and national tests of Obstetrics and gynecology group were found in 53% and 54% of the test questions, respectively. The analysis revealed that 30% and 37% of the questions were easy and 17% and 9% were difficult. (p=0.08)(Table 3)
In pediatric group proper Dif. Indices were obtained in 34% and 43%, easy indices were obtained in 60% and 51% and Difficult indices in 6% and 7% in local and national tests, respectively.(P=0.28) (table 3)
In internal medicine group proper indices were obtained in 40% and 42%, easy indices in 42% and 46% and difficult ones in 18% and 12% (P= 0.34)(table3)
Generally, in local and national tests given to all 3 groups, most of the questions were reject able. Few optimal questions were observed.
In Obstetrics and gynecology group 76% of questions in both local and national tests were "must omitted" and just zero and 1% were optimal. "must omitted" questions in pediatric group were 81% and 79% and optimal ones 1% and zero, respectively. In internal medicine group 91% and 85% of questions were "must omitted" and 1% was optimal in the both tests.
Question in the local and national tests were statistically similar questions in the 3 residency groups under the study. (Table 4)
The most common causes making the questions to be considered "must omitted" in local and national tests of all 3 groups (Obstetrics and gynecology, pediatrics and internal medicine) were negative, zero or less than 0.2 Discrimination indices (table 5).

Discussion

A study in the university of North Dakota, in order to improve medical school essay questions statistics was used regarding item difficulty and

Table 1. Comparison of questions in local and national tests regarding cognitive domain level in 3 residency groups of Hamadan University of Medical Sciences in 2004

Cognitive domain level	Local	National	Significance level
Obstetrics and gynecology group: "recall" "application of knowledge"	107(72) 41(28)	107(71) 43(29)	P=0.85
Pediatrics group "recall" "application of knowledge"	104(72) 40(28)	53(35) 97(65)	P<0.001
Internal medicine group "recall" "application of knowledge"	76(51) 74(49)	28(19) 122(81)	P<0.001
Numbers indicate frequency (%)			

Table 2. Comparison of questions in local and national tests regarding discrimination indices in 3 residency groups of Hamadan University of Medical Sciences in 2004

Discrimination index	Local	National	Significance level
Obstetrics and gynecology groups: 0.7-1 0.5-<0.7 0.2-<0.5 <0.2 or 0 or negative	4(3) 17(11) 54(36.5) 73(49.5)	8(5) 13(9) 47(31) 82(55)	P=0.413
Pediatrics group: 0.7-1 0.5-<0.7 0.2-<0.5 <0.2 or 0 or negative	5(3.5) 18(12.5) 35(24) 86(60)	1(1) 23(15) 45(30) 81(54)	P=0.217
Internal medicine group 0.7-1 0.5-<0.7 0.2-<0.5 <0.2 or 0 or negative	1(1) 9(6) 32(21) 108(72)	1(1) 12(8) 24(16) 113(75)	P=0.652
Numbers indicate frequency (%)			

Table 3. Comparison of questions in local and national tests based on difficulty indices in 3 residency groups of Hamadan university of medical sciences in 2004

Difficulty index	Local	National	Significance level
Obstetrics and gynecology group: Proper Easy Difficult	78(53) 45(30) 25(17)	81(54) 56(37) 13(9)	P=0.08
Pediatrics group: Proper Easy Difficult	49(34) 86(60) 9(6)	64(42) 76(51) 10(7)	P=0.28
Internal medicine group: Proper Easy Difficult	60(40) 63(42) 27(18)	63(42) 69(46) 18(12)	P=0.34
Difficulty index of 30-70= proper > 70= easy <30= different			

Table 4. Comparison of questions in local and national tests based on general quality in 3 residency groups of Hamadan University of medical sciences in 2004

Quality	Local	National	Significance level
Obstetrics and gynecology group: Optimal Proper Acceptable "must omitted"	0(0) 4(3) 31(21) 113(76)	1(1) 7(5) 28(19) 114(76)	P=0.58
Pediatrics group: Optimal Proper Acceptable "must omitted"	1(1) 4(3) 22(15) 117(81)	0(0) 1(1) 30(20) 119(79)	P=0.27
Internal medicine group: Optimal Proper Acceptable "must omitted"	1(1) 1(1) 13(9) 136(91)	1(1) 1(1) 21(14) 128(85)	P=0.6

Table 5. Frequency of "must omitted" questions by the cause in local and national tests in 3 residency groups of Hamadan university of medical sciences in 2004.

cause of rejection	Local	National	Significance level
Obstetrics and gynecology group:			
A*	73(64.5)	79(72)	
B**	16(14.5)	13(11)	P=0.893
C***	24(21)	19(17)	
Pediatrics group:			
A	86(74)	81(68)	
B	26(22)	31(26)	P=0.925
C	5(4)	7(6)	
Internal medicine group:			
A	108(80)	113(88)	
B	14(10)	13(10)	P=0.063
C	14(10)	2(2)	

**A: negative or zero or less than 0.2 DIs*

***B:DIs between 0.2 –0.5 coexisting with PI of more than 80(very easy) or less than 20(very difficult)*

****C: DIs between 0.2-0.5 and PIs between 20-80 coexisting with just "recall" level level.*

discrimination value. Results indicate that statistical analysis of tests provides valuable feed back for improvement of essay questions (6)

Residents who are training for a highly intellectual clinical practice should be evaluated application of knowledge in their academic achievement tests. Just "recall" level questions in 71-72% of obstetrics and gynecology questions reveals a serious problem requiring attention and a change in the evaluation system.

The same problem is observed in pediatrics and internal medicine tests, and it is more prominent in local than national tests.

Quality improvement can be achieved through determination of a certain percentage of "recall" questions and it should be controlled using a check list before the final approval of the test.

Designing MCQS which demand more than recall from students is more difficult than asking factual questions. MCQs can test much more than recall (5).

Discrimination and difficulty indices are two significant indicator of a question specially in higher academic tests. All the three studied groups were statistically similar in local and national tests given in early 2004 regarding.

DIs and Dif indices and the cognitive level mix of the tests were not acceptable. Few questions were optimal. In contrast, "must omitted" questions were 76% in Obstetrics and gynecology, 85-91% in internal medicine and 79-81% in pediatrics. These high percentages of "must omitted" questions confirm the problem and the necessity to change the evaluation system.

In a study in medical faculty of Mashad university, analyzing 3000 MCQs taken in first semester of 2000- 2001, item difficulty of more than 50% of questions were greater than 0.7. In this study, only 16.5% of MCQs' difficulty index ranged between 0.5 and 0.6 and 52.2% of MCQs had inappropriate item difficulty index. Findings of this study indicate that only 33% of studied MCQs have desirable or acceptable item difficulty and Discrimination indices both and 34.9% of those have no desirable or acceptable item difficulty index and discrimination index (7).

The ratio of questions of application level should be increased with proper difficulty indices (30-70) and discrimination indices (0.7 or more).

The first step could be concentration on discrimination index, because the most prevalent causes of considering a question as "must omitted" were negative or zero values of discrimination indices (42-74%). Among the most useful, approved methods are challenging suggested questions by presumptive test participants (screening the questions) and using question banks.

Reference

1. Seif A. Educational measurement and Evaluation methods (Persian translation). 2th ed. Tehran: Douran Publication; 1997 .

2. Mavis BE, Cole BL, Hoppe RB. A survey of student assessment in U.S. medical school: The

balance of breadth versus fidelity. Teach learn Med. 2001; 13(2): 74-9.
3. Mavis BE, cole BL, Hoppe RB. A survey of information sources used for progress decisions about medical students .2000;5:9. Available from: URL http://www.med-ed-online.org.
4. Lavine AR. Test item difficulty related to item discrimination and content. An approach to improve test construction. 8th Annual meeting of the international association of medical science educators. 2004. Louisiana U.S.A.
5. Higgins E, Tatham L. Assessing by multiple choice questions (MCQ) tests 2003. Available from:http://www.ukcle.ac.uk/resources/trns/mcqs/index.htm.
6. Edward G. Simanton. Improvement of medical school essay questions using statistics. 8th annual meeting of the international association of medical science educators. Louisiana USA 2004.
7. Tabatabaee M, Bahreyni Toosi M, Derakhshan A, Khayeh Dalloee M, Gholami H. Analytic assessment of multiple choice tests, JME. 2003;2: 87-91.