Res Dev Med Educ, 2014, 3(1), 31-36
doi:10.5681/rdme.2014.008
http://journals.tbzmed.ac.ir/rdme

Original Research

Comparison of Multiple-Choice Questions in Quality Parameters of Pediatric Residency Tests between the Pre-Board Examination of Tabriz University of Medical Sciences and National Board Examination in 2007 and 2011

Mohammad Barzegar[1,2]*, Nemat Bilan[1,2], Mohammad Hassan Karegar Maher[2], Siamak Shiva[2], Manizheh Sayyah Melli[3], Aydin Tabrizi[2]

[1] *Medical Education Research Center, Tabriz University of Medical Sciences, Tabriz, Iran*
[2] *Pediatric Health Research Center, Tabriz University of Medical Sciences, Tabriz, Iran*
[3] *Women's Reproductive Health Research Center, Tabriz University of Medical Sciences, Tabriz, Iran*

Article info

Article History:
Received: 4 Jan 2014
Accepted: 21 Feb 2014
ePublished: 29 May 2014

Keywords:
Multiple Choice Questions,
Structural flaws,
Taxonomy,
Pediatric residency

Abstract

Introduction: The objective of this study was to compare Multiple-Choice Questions (MCQs) quality parameters of pediatric residency tests between the pre-board examinations of Tabriz University of Medical Science (TUMS), Tabriz, Iran and the national board examination in 2007 and 2011.

Methods: In this cross-sectional study, we evaluated the format of 300 MCQs in the pre-board examination of TUMS and the format of 300 MCQs of the national board examination in pediatric residency. Individual MCQs were evaluated for content budgeting according to the Nelson pediatric residency reference textbook, taxonomy levels (Bloom's levels I, II and III) and following structural principles (based on Millman checklist). Data were analyzed by SPSS (version 18) software.

Results: We find more consistent content budgeting in the national board MCQ examinations. Forty one percent of pre-board MCQ examinations and 72% of national board MCQ examinations were Bloom's taxonomy levels II -III (P=0.000). We found correct structural principles in 69.2% and 76.2% of pre-board and national board MCQs examinations, respectively (P=0.05). 30.7% and 22.5% of pre-board and national board MCQs examinations were negative stem, respectively (P=0.025). Most of the negative stem MCQs were Bloom's taxonomy level I questions.

Conclusion: Pediatric residency pre-board MCQ examinations of TUMS were of a significantly lower level of learning (taxonomy level I) compared to the national board MCQ examinations. To prevent low quality development of internal university examinations, monitoring of these exams is recommended.

Introduction

Assessment of students forms one of the most important parts of any education program. Assessment involves a systematic process of collection, analysis, and interpretation of data in order to determine if educational goals have been or are being realized, and if so, to what level.[1] The assessment process should produce an appropriate picture of the academic progress of every student at different time scales and identify problems and deficiencies within their education. When results are unsatisfactory, it may indicate poor effort on the part of the student, but it may be due to failure in planning, teaching, or improper assessment.[1-4] There are numerous methods of assessment, which are determined based on the purpose of the assessment. Board certification and promotion examinations in clinical disciplines of medical residency courses are held annually, which are certification-type assessments used to rate and determine promotion to higher grade or to award a Science (Medical) Degree in order to protect the society against incompetent practice.

Currently, written multiple-choice tests are the most common objective examinations for this purpose. Multiple-choice tests can evaluate a wide range of students' knowledge in a short period of time. They are the best objective tests in terms of uniformity of questions, possibility of blind guessing (compared to true or false tests), and convenience of scoring. Multiple-choice questions (MCQs) use one question or problem as the base or stem question, and 4-5 answers as options, where

***Corresponding authors:** Mohammad Barzegar, Email: mm_barzegar@yahoo.com

only one is correct. Incorrect answers, as possible answers to the question are also offered as distractors beside the correct answer. If MCQs are prepared carefully and test development principles are observed, the asessment can properly distinguish between highly and low competent students.[1-7] Preparation of good MCQs for an experienced question designer requires at least one hour.[8] Obviously, this process takes longer for inexperienced people. In recent years, MCQs -designing workshops have been held for nearly all faculty members of Tabriz University of Medical Sciences (TUMS) for proper design of MCQs. Since 2006, designing and holding examinations for the promotion of residents have been assigned to Type 1 universities of medical sciences in order to increase the participation of all faculty members in teaching groups of the country's universities of medical sciences, in accordance with the Medical and Specialized Education Council program, which is in line with progress toward autonomy of universities, and extended delegation of educational authorities. Therefore, it is necessary to evaluate this process (designing of questions and holding promotion examinations by universities).

Although several studies have been conducted on the analysis of MCQs in Iran, they mainly aimed to provide feedback for question designers.[9-11] This study aimed to compare some quality indicators of MCQs (content budgeting, taxonomy distribution and adhering to structural principles of standard) in pediatric residency promotion examinations (2007 and 2011) held at TUMS, with board certification questions in this discipline during mentioned years. Special consideration was also given for improved quality management and monitoring of university internal examinations.

Materials and Methods

This was a cross-sectional descriptive study of 600 questions in the form of written examinations for promotion, including 150 questions for each year (2007 and 2011) and board certification examination, including 150 questions for each year (2007 and 2011), in pediatric residency. Study questions were selected by census. Questions related to promotion and board certification examinations were evaluated for each year in terms of taxonomy by the project executive and two expert pediatric residency education professors (one of whom was a member of the national pediatric board certification examination). To homogenize professor perceptions of taxonomy, a guideline was sent to each, so they could divide questions in terms of taxonomy, as follows:[2,12]

Taxonomy I - ability to remember facts (memory),
Taxonomy II - ability to interpret data
Taxonomy III - ability to solve a new problem

When consensus was not reached on taxonomy of a question, a fourth expert pediatric residency education professor determined taxonomy, and his opinion was accepted. All test questions were adjusted by the project executive according to the Millman checklist, and were assessed as follows (Table 1):

Unfortunately, case 1, "Has the purpose been listed in the question's information form?", was not assessed due to the unavailability its information form, and case 4, "Distracters should be written, so as to draw attention of unknowledgeable examinees," was not assessed due to unavailability of answers and scores of examinees.

Data collected and analyzed by SPSS-18 software. Given the qualitative nature of variables that were compared between promotion and board certification tests, a Chi-square test was used to investigate the significance of difference between variables of the two tests. $P<0.05$ was considered significant.

Results

No question was designed from the topics of pediatric surgery and disorders of the eye in the 2007 promotion examination, and no question was designed for pediatric surgery or skin disorders in the 2011 promotion questions, which means that board examinations had good content coverage.

Table 1. Checklist for Reviewing Multiple-Choice Items

Content	yes	no
1. Has the item been constructed to assess a single written objective?		
2. Does the stem include the least amount of information necessary for understanding the question and selecting the correct answer?		
3. Are the alternatives free from clues as to which response is correct?		
4. Distracters should be written, so as to draw attention of unknowledgeable examinees.		
5 . Does highlighted negative words in negative stem questions?		
6.Is the key the same length and level of detail as the distracters		
7. Have the alternatives "all of the above" and "none of the above" been avoided?		
8 . Are the alternatives homogeneous in content?		
9. Are the grammar, punctuation, and spelling correct?		
10. Can you understand what is being asked without reading the options?		
11-Not be used two opposite option which one of them is correct		
12.Avoiding from repeating content in options		

The Kappa consistency coefficient between evaluating professors was 0.59 for determining taxonomy level, which was within the acceptable range. Since agreement was not reached by 3 professors in 10 questions, taxonomy was determined by the fourth professor. Taxonomical distribution of questions in different tests is presented in Table 2.

Medical Science University and the national board examination for pediatric residency in 2007 and 2011

In terms of taxonomy, 26.6% of questions in 2007 promotion examination, and 60.6% of questions in 2007 board certification examination were in taxonomy levels II-III (*P*=0.000). In 2011 promotion examination, 54.2%, and in 2011 board certification examination, 69.3% of questions were in taxonomy levels II-III (*P*=0.000). Assessment of the trend of tests showed increasing percentage of taxonomy II-III in every type of test. A comparison between 2007 and 2011 promotion tests was significant (*P*=0.000). The difference between 2007 and 2011 board certification tests was not significant (*P*=0.116). Overall, in both of years,41 and 65% of questions were within taxonomy levels II-III for promotion and board certification tests, respectively (*P*=0.000).

A comparison of the numbers and percentages of questions in taxonomy level II-III in some topics of the promotion examinations of TUMS and national board certification for pediatric residency in the study years are presented in Table 3.

Examination of Tabriz Medical Science University and the national board examination for pediatric residency in 2007 and 2011

In terms of compliance with Millman's structural principles in 4-option questions, the 2007 and 2011 promotion and board certification examinations of pediatric residency are compared in Table 4.

An improvement has been achieved in observing structural principles in 2011 written promotion examination (70.7%) compared to 2007 (67.8%) (*P*=0.58). This difference was significant in the written board certification examinations in the above years (*P*=0.013). In the combined 2007 and 2011 promotion tests, 69.2% of promotion questions and 76.2% of board examination questions were without structural problems (*P*=0.05). The highest frequency of structural problems was observed in 27.5% of 2007 promotion questions, 25% of 2007 board certification questions, 21% of 2011 promotion questions, and 12.2% of 2011 board certification questions, which were associated with 10 Millman list cases of "Answering without attention to options", and 2 Millman list cases of "expressing great part of information in the stem of question".

A comparison of numbers and percentages of questions with negative stems in written promotion examinations of TUMS and national board certification for pediatric residency in the study years are presented in Table 5.

Table 2. Taxonomy distribution of questions of the pre board examination of Tabriz

Level of taxonomy	I	II	III	Total
Test	Number (%)	Number (%)	Number (%)	
Preboard 2007	110 (73.3)	34 (22.6)	6 (4)	150 (100)
Board 2007	59 (39.3)	71 (47.3)	20 (13.3)	150 (100)
Preboard 2011	69 (46)	64 (42.6)	17 (11.3)	150 (100)
Board 2011	47 (31.3)	83 (55.3)	20 (14)	150 (100)
Total	285 (47.5)	252 (42)	63 (10.5)	600 (100)

Table 3. Taxonomy (II-III) distribution of questions in some topics of the pre-board

Number of questions in topic	Pre-board		Board		*P*. value
Topics	Total	Number (%) II-III	Total	Number (%) II-III	
Growth, Development And Nutrition	24	7 (29.3)	26	9 (34.6)	24
Pediatric Drug Therapy& The Acutely Ill child and Environmental Health Hazards	17	7 (41.2)	18	14 (77.8)	17
Fetus and the Neonatal infant	43	14 (32.6)	38	28 (73.7)	43
Infectious Diseases	44	15 (34.1)	46	32 (69.5)	44
Cardiovascular System	16	9 (56.3)	14	10 (71.4)	16
Nephrology and Urologic disorders	18	10(55.6)	14	11 (78.6)	18
Endocrine System and Metabolic diseases	21	12 (57.1)	16	9 (56.3)	21
Nervous system and Psychologic Disorders	21	12 (57.1)	22	15 (68.3)	21

The difference between negative-stem promotion questions for the 2007 and 2011 examinations was significant (*P*=0.045). The difference in national board certification in 2007 and 2011 was insignificant (*P*=0.081). The correlation between questions' taxonomy and negative stem was significant in all promotion and board certification tests (*P*=0.000). 68.6% of negative-stem questions in the combined 2007 and 2011 promotion tests, and 52.2% of negative-stem questions in the combined 2007 and 2011 board certification tests had been designed in taxonomy I.

Table 4. Comparison of compliance of Millman's structural principle in the pre-board examination of Tabriz Medical Science University and national board examination pediatric residency in 2007 and 2011

Written examination	Pre-board		Board		*P*. Value
	no structural problem	structural problem	no structural problem	structural problem	
	Number (%)	Number (%)	Number (%)	Number (%)	
2007	101 (67.8)	49 (32.2)	105 (70)	35 (30)	P=0.68
2011	106 (70.7)	44 (29.3)	124 (82.7)	26 (17.3)	P=0.018

Table 5. Comparison of questions with negative stems in pre-board examination of Tabriz medical science university and national board examination pediatric residency in 2007 and 2011

Written examination	Pre-board		Board		P. value
	Negative stem Number (%)	Positive stem Number (%)	Negetive stem Number (%)	Positive stem Number (%)	
2007	54 (36)	106 (64)	40(26.8)	110 (72.2)	P=0.46
2011	38 (25.3)	112 (74.7)	28(18.6)	122 (81.3)	P=0.146

Discussion

Changing of curriculum or teaching methods without changing assessment methods will not produce desirable results. Furthermore, changing the assessment system, even without implementing changes in curriculum, leaves a deeper effect on the nature of learning, compared to changes in curriculum without changing the assessment system.[1-4]

In analysis of a question, both qualitative and quantitative aspects are considered. In qualitative terms, the form of the question, preparation method, care applied in the question's text, right and wrong options, and taxonomy of the question are evaluated. In quantitative terms, the degree of difficulty and distinctive ability of each question and analysis of distracters are examined.[1-4,7,13-14]

A good test is one which can best reflect all training aims and contents of curriculum design. When curriculum is more detailed, the test designer can choose questions which best indicate the contents and aims of a curriculum instead of placing all possible questions of all contents and aims in a test using a questions features table (test blueprint).[1-4,8] Regarding course content evaluation, it should be mentioned that though questions' percentage distribution in different topics of reference books were different for different years of examination, there were no fixed criteria for questions' content budgeting, even for the board examination. However, board examination questions should cover a perfect sample of course content. For example, no question was designed for pediatric surgery, eye, ear, throat and nose in promotion questions of 2007.

One of the most important consequences of designing no questions from some topics is the selective omission of issues by the resident. Observing an appropriate ratio of questions of different issues (content budgeting), and being faithful to that ratio during the test designing procedure and continuing it for subsequent years is one characteristic of a fair test. This issue was not observed in studied exams. The most appropriate condition of budgeting is to determine learning necessities based on each course curriculum and devoting an appropriate number of questions according to this criteria. One major problem in determining the percentage of questions in different topics is the nature of some theoretical questions that can be divided into different disciplines. For example, a question covering an asthma issue can also be related to allergy or respiratory system disease.

Considering the importance of the medical code of ethics, at least two questions are considered in this case during all board exams held in recent years whose scores are those other than the test scores level. These questions are considered a bonus for those who answer correctly. It is suggested that this procedure would be repeated in promotion exams.

According to the results obtained in the present study, 73.3% of 2007 promotion test questions in pediatric residency examinations were in taxonomy level I. Although there are no consistent standards for the percentage distribution of questions for different taxonomy levels, it is recommended that this figure be less than 50%. In a study by Mohagheghi et al., in a taxonomical assessment of questions in the 2007 written board examinations in 25 clinical specialist disciplines, 38.7±18.9% of questions had been designed in taxonomy levels II-III; this figure was 45±19.3% in the

2008 board and 56±15.51 percent in the 2009 board, which indicates a growing trend of question design at higher taxonomy levels.[15] In examining 2400 questions relating to residency examinations of Gundi Shapoor University of Medical Sciences in 2007, only 28% of questions had been designed at taxonomy levels II-III.[9] Unfortunately, this figure is even less in non-medical internal examinations. In a study by Sanagu et al., less than 5% of 523 nursing examination questions were within taxonomy levels II-III.[10] It appears the problem of overcoming memory-based questions (Taxonomy I) in examinations is a deeply-rooted problem in the medical education system that extends from internal university examinations to national examinations. Some studies show that such a problem not only taints test validity, but also pushes students to superficial learning and memorizing. Unfortunately, one of the drawbacks of multiple-choice tests is their excessive use of low level learning.[1-7] There was a significant difference in taxonomy levels in written examination questions in promotion and board certification in pediatric residency.

In total, 41% of questions in the 2007 and 2011 combined written promotion examinations were designed in taxonomies II-III, and this figure for the national board certification examination in those years was 65% (P=0.000). Luckily, significant progress occurred in designing of questions at deeper learning levels (taxonomies II-III) in the 2011 promotion examinations compared to those of 2010 (P=0.000), which indicates an improved trend in the question designing skills of faculty members. Given the important role that these examination results play in determining the competence, capability, and qualification of students for graduation or promotion to higher level, if questions are designed in lower cognitive levels, then the test validity will be questionable. In clinical education, acquiring the necessary capabilities in dealing with the patient and the disease are essential in this education course, and student's ability to deal with a patient and his competence in the required skills should be evaluated. Whether or not students can appropriately recognize and adopt the right decision about patients with the necessary competence cannot be made certain merely based on memory based questions (Taxonomy I). In the long term, students' learning activities and what they learn is determined by the type of examinations they must pass. Therefore, if teachers use assessments that merely require memory, then students will be encouraged to simply memorize the subjects. Conversely, if the examination is so designed that the student has to consider principles, interpret information, or solve a problem, then it will create the tendency in student to learn such points, so he can pass the exam with flying colors. The significant differences in the level of questions in university examinations will lead to superficial study by the residents and their reduced success rate in board certification exams.

Considering that the questions' taxonomy level distribution was different in different subgroups (subspecialty), providing feedback to test designers who mostly had taxonomy I and providing needed requirements to promote this skill is necessary. It was shown in several studies that providing feedback to test designers had a positive effect on promoting question quality.[16,17]

According to the Millman list, nearly 46% of 2010 promotion questions had no structural problems. This figure in the 2011 promotion exam rose to 64%, producing a significant difference. Results of the study by Mohagheghi et al., on the board certification exam between 2007 and 2009 in 25 specialized clinical disciplines showed that 57.5±15.1% of questions in the 2007 board certification test and 63.8±15.5% of questions in the 2008 board certification test, and 60.6±18.9% of board exams had no structural problems, based on Millman's structural principles in multiple-choice questions.[15] A study by Shokornia et al. as well as other studies reported different levels of structural problems in multiple-choice tests.[9-11] A study in the U.S. reported 46% structural problems in multiple-choice questions,[18] and two similar studies in 2006 and 2008 in Hong Kong also showed 46.2% and 47.3% of multiple-choice questions used in the assessment of students' problems had structural problems and were at low cognitive levels.[19,20] Unfortunately, a significant proportion of promotion questions contained structural problems. The most common structural problems (according to Millman principle) that this study showed were the following: "answering questions without reading the options" and "A great part of information not being in the stem of question", which is similar to other studies,[9-11,15] and it may be due to the established old habits in designing questions.

One of the difficulties in preparing multiple-choice questions is that sometimes, it is difficult to prepare incorrect distracters that appear right. In such cases, if it is easier to prepare correct options, then negative-stem type questions in which all options except one are right can be used. Unfortunately, a majority of the negative-stem questions in the examinations studied had been designed at taxonomy level I, and the correlation between questions' taxonomy and negative-stem was significant in all promotion and board certification tests. However, according to ministerial guidelines for question designers, if a negative word is underlined, it is structurally considered flawless. Given the undesirable effect of negative-stem in the taxonomy of questions, it is recommended that a negative-stem question be structurally considered undesirable. Some studies indicate that use of negative verbs in the stem of the question will lead to confusion of the person being tested. In these circumstances, the respondents have to change the question from negative to positive in their mind and then find the right answer. Positive-stem multiple choice questions, compared to negative-stem ones, possess higher levels of validity and reliability and better assess students' academic performance.[21-23]

There is no doubt that the delegation of promotion examinations to universities of medical sciences is a very valuable approach. To prevent lowered quality of internal university examinations, monitoring exams by universities with supervision of Medical and Specialty Education

Council Office will lead to reduction in deficiencies.

Conclusion

Pediatric residency promotion examination questions of Tabriz University of Medical Sciences had been designed at significantly lower learning level (taxonomy I) compared to the national board certification examination. Planning for the empowerment of question designers in the area of test-making can help correct and improve the preparation and design of suitable multiple-choice questions.

Study limitations

For quantitative analysis, students' scores and their answer sheets are required. Unfortunately, in this study, this information was not available. Hence, quantitative analysis of tests was not possible. Due to the unavailability of board certification question forms in which question design purpose is identified examining the budgeting of questions in terms of goals of the educational course was not possible, which was the main weakness of this study which should be addressed in future studies.

Acknowledgements

This article was part of the research project approved by the Medical Education Research Center of Tabriz University of Medical Sciences, and the corresponding author's thesis for a Master's Degree.

Competing interests

The authors declare that there is no conflict of interests.

References

1. Seif A. [Educational measurement, assessment and evaluation]. 5th ed. Tehran: Doran; 2008.
2. Guilbert JJ. Education handbook for health personnel. 7th ed. Geneva: World Health Organization;1998.
3. Bend David MF. Principles of assessment. In: Harden RM, Dent JA. A practical guide for medical education. 2nd ed. London: Elsevier;2005.
4. Zolfaghari B, Asadollahi GH. Academic achievement tests in medical sciences. Isfahan: Isfahan University of Medical Sciences; 2000.
5. Burton SJ, Sudweeks RR, Merrill PF, Wood B. How to prepare better multiple choice tests items: guideline for university faculty. Birgham: Young University Testing Services;1991.
6. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education 2002;15:309-34.
7. National Board of Medical Education. Constructing written test questions for the basic and clinical sciences[internet]. Philadelphia: National Board of Medical Education;2002. Available from: http://www.mf.uni-mb.si/gradiva/ang/SBA_MCQ_NBME.pdf
8. Farley JK. The multiple-choice test: developing the test blueprint. Nurse Educator 1989;14(5):3-5.
9. Shakournia AH, Mozaffari AR, Khosravi Broujeni A. [Survey on structural of MCQs of residency exam in AJUMS]. Judishapur Scientific Medical Journal 2010;8:491-502.
10. Sanagoo A, Jouybari L, GhanbariGorji M. [Quantitative and qualitative analysis of academic achievement tests in Golestan University of Medical Sciences]. Research in Medical Education 2010;2:24-32.
11. PourmirzaKalhori R, Roshanpour F, Rezaei M, Naderipour A. [Knowledge improvement effect on results of multiple choice questions in residency exams analysis (2009)]. Journal of Kermansha University of Medical Sciences 2010;15:112-8.
12. Buckwalter JA, Schumacher R, Albright JP, Cooper RR. Use of an educational taxonomy for evaluation of cognitive performance. J Med Educ 1981;56:115-21.
13. Collins J. Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules. Radiographics 2006;26: 543-51.
14. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. 3rd ed. Philadelphia: National Board of Medical Examiners;1998.
15. Mohaggegi MA , VahidShahi K, SHakeri S, Saburi M, Razavi M , Mohammadi M, et al. [Comparison some aspect of quality of MCQs Board Examination: 2007-2009[internet]]. Available from: cgme.behdasht.gov.ir/uploads/264_781_N4
16. Shaban M, Ramazani Badr F. [Effect of test item analysis on summative exams on quantity of test designing]. HAYAT 2007; 13:5-15.
17. Danish KF, Ahmad Khan R. Role of effective feed back in Multiple Choice Questions (MCQs) designing for faculty development. Journal of Rawalpindi Medical College 2010;14 :98-100.
18. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test item on achievement examinations in medical education. Adv Health Sci Educ Theory Pract 2005;10:133-43.
19. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Educ Today 2006;26:662-71.
20. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 2008;42(2):198-206.
21. Harasym PH, Price PG, Brant R, Violato C, Lorscheider FL. Evaluation of negation in stems of multiple-choice items. Eval Health Prof 1992;15:198-220.
22. Harasym PH, Doran ML, Brant R, Lorscheider FL. Negation in stems of single response multiple-choice items: an over estimation of student ability. Eval Health Prof 1993;16:342-57.
23. Dudycha AL, Carpenter JB. Effects of item format on item discrimination and difficulty. J Appl Psychol 1973;58:116-21.