Original Article

Quality of design and analysis of pre-internship exams for medical students of Hamadan University of Medical Sciences in the years 2018 to 2020

Masoumeh Rostami-Moez [1], Mahdi Biglarkhani [2], Azam Meyari [*2]

[1] Research Center for Health Sciences and Education Development Center, Hamadan University of Medical Sciences, Hamadan, Iran.
[2] Persian Medicine Department, Medical School, Hamadan university of Medical Sciences, Hamadan, Iran.

Article Info

Article history:
Received 19 Nov 2021
Accepted 09 Mar 2022
Published 18 Mar 2022
Keywords:
Pre-internship exam
Test analysis
Taxonomy
Question structure
Multiple Choice Questions

***Corresponding author:**
Azam Meyari, Persian Medicine Department, Medical School, Hamadan University of Medical Sciences, Hamadan, Iran.
Email: a.meyari@umsha.ac.ir

Abstract

Background & Objective: Evaluation is a major step in the educational process that measures learners' abilities and test analysis is a tool to measure its effectiveness. The pre-internship exam is a test by which qualification of medical assessed before entering the internship stage. Therefore, its analysis can indicate its evaluation of the competencies of medical students. In this study, the quantitative and qualitative indices of pre-internship exams of medical students of Hamadan University of Medical Sciences were studied.

Materials & Methods: In this documentary research, all pre-internship questions in September 2018 to 2020, were assessed in terms of taxonomy, structural defects, and difficulty, discrimination indexes. The taxonomy of the questions was evaluated based on Bloom's classification, the structure of the questions based on the Millman checklist, and the index of difficulty and discrimination using the Excel forms available at the university. Data analysis was performed using SPSS 20 software.

Results: In 2018, 56% of the questions and in 2019 and 2020, 64% of the questions were designed at the level of taxonomy 2 and 3. More than 90% of the questions haven't any structural defects. The difficulty index of the questions was on average at the appropriate level, but the discrimination index of the questions was moderate in 2018 and 2020 and weak in 2019.

Conclusion: In attention to the amount of questions with a negative discrimination index, the review of questions, options and the key questions after designing the exam is suggested.

Copyright © 2021, This is an original open-access article distributed under the terms of the Creative Commons Attribution-noncommercial 4.0 International License which permit copy and redistribution of the material just in noncommercial usages with proper citation

Introduction

Assessment is a basic step in the learning process of learning. It measures the academic performance of a learner during a training course. One of the methods to assess the training is to use multiple-choice questions. Analysis of the Tests could actually be defined as an educational tool to reveal the learners' competence as well as the gap between educational goals and the amount of learning. There are different methods to assess the level of knowledge for learners and to be ensured the achievement of educational goals (1).

Multiple choice questions are among the types of written questions that are commonly used in course activities based on the theoretical content of the course (2-4). Although multiple-choice questions generally assess low levels of knowledge, if properly designed and coincident with learning objectives, they can be applied to assess high levels of knowledge, i.e., comprehension, analysis, application, and problem solution. In addition to knowledge, good writing skills are required to write good multiple-choice questions (6).

To complete the training process during the course, it is necessary to study and analyze the quality of the questions. Therefore, test analysis is an integral part of course evaluation. In other words, test evaluation is a dynamic process that aims to improve questions and teaching (6) Student evaluation is one of the most important tasks that a professor faces, however, the

[Downloaded from edujournal.zums.ac.ir on 2026-11-09]

[DOI: 10.66224/edcj.14.44.36]

quality of many assessment methods and exams is less than desirable. If the evaluation is based on scientific principles and standards, it can be considered as the most important pillar of education and the most effective factor for improving the quality of learning (7).

The evaluation theory of the classical test is based on an equation in which the observed score is hypothetically composed of a true score and an error score. In this model, the real score of each person is fixed and the measurement errors will be random. In classical theory, criteria including difficulty, mean, variance, validity, and validity of the test are estimated and calculated (8) Also, the analysis of questions is done in two ways, qualitatively and quantitatively. Quantitative analysis of questions examines the three components of level of difficulty or ease, the power of question clarification, and deviant options (9). Examining the structure of the questions using the Millman checklist is included in the qualitative examination of the questions and refers to the percentage of questions without any structural problems (3,4) Millman's most important goals in designing multiple-choice questions include putting more question information in the body of the question, using the question to gauge a learning goal, using fluent and clear words, not using negative words or phrases in the question options, or highlighting them, not using answers with the words "all options" or "none", not using contradictory options, independence of questions from each other, the equivalence of questions in length and vocabulary structures, not using repetitive words, no spelling mistakes and arrangement of the options (10).

In addition to this qualitative analysis, it refers to the percentage of questions that are classified based on educational purpose using Bloom's taxonomy and includes knowledge level classification (I), understanding and concept (II), and application (III) (11, 3). The use of valid multiple-choice questions at high taxonomic levels helps medical students to increase their conceptual understanding and to improve cognitive reasoning and methodological knowledge development for the complexities of clinical practice (11).

A pre-internship test is a test by which the competence of medical students is measured before entering the internship stage to enter the next stage if the necessary conditions are met, which is the decision-making stage for patients. Given the special importance of this test to measure the knowledge of future physicians and ultimately improve the health of the community, it is clear that the more standard and scientifically-based questions are designed, the better the tool will be to differentiate unscientific students from scientifically qualified students. Numerous studies have been published on the success or failure of pre-internship exams and various factors affecting its results have been reported. Success or failure of medical students in factors such as GPA of diploma and GPA of basic sciences, physiopathology and internship, the score of pre-internship exams, probation history, admission quota, gender, duration of the study and student employment have been studied. The results are in some cases contradictory (12). The general evaluation of the multiple-choice pre-internship multiple-choice test questions helps to improve the level of tests based on scientific evidence, and as a result, the improvement of the tests is used to improve the educational level and assess students' learning (13, 14). Therefore, due to the importance of pre-internship exams at the beginning of the internship and the need for its analysis (12). The study was conducted to evaluate the multiple-choice pre-internship exam questions for medical students of Hamadan University of Medical Sciences in the last three courses (2018-20) to use the results to improve the educational level and assess students' learning.

Materials and Methods

The study is documentary research. This study aimed to investigate pre-internship exams in medicine in the years 2018 to 2020. The assessment tool was performed by reviewing the pre-internship multiple-

[Downloaded from edujournal.zums.ac.ir on 2026-11-09]

[DOI: 10.66224/edcj.14.44.36]

choice exam held in the examination center of Hamadan University of Medical Sciences (a large subset of the 3-country health zone). In this study, all pre-internship exams in medicine were held in the examination center of Hamadan University of Medical Sciences based on classical test theory (8) and Millman checklist (10) are analyzed in the last three periods (September 2018, September 2019, and September 2020). The tests were evaluated in terms of quantitative indicators (difficulty index, discrimination index, and test validity) and qualitative indicators (percentage of compliance with structural design principles and percentage of question design in levels I, II, and III taxonomy). In this study, to have challenges and differences in differentiating and determining the level of taxonomy II and III, the questions were divided into questions with high taxonomy II) and III (and questions with low taxonomy) (I). The structural problems of the questions were assessed using the Millman standard checklist. This checklist contains 18 questions and the answer is scored based on yes and no. Each question was reviewed by a physician and expert in the field of medical education and according to the Millman checklist, all items in the checklist were reviewed and in case of non-compliance in the checklist, each question was noted. Also, at the end of the taxonomy, the question was identified by the same expert in the field of medical education and was noted at the end of the checklist. To establish inter-rater reliability, coordination was established during the meetings to achieve these goals. In quantitative indicators, the level of difficulty indicates the percentage of people who answered the question correctly, and by dividing the number of people who answered the test correctly by the number of people who took the test, it is calculated and less than 0.3 difficult test, and is classified between 0.3 to 0.7 suitable and more than 0.7 easy tests (3).

Distinctive power or discernment indicates the power of discernment of people with high ability (high score) and low ability (low score). The range of discrimination index of the question is in the range of+-1. Zero range means that the question lacks discernment. Ranges less than zero, 0-0.12, 0.13-0.39, and 0.40 and more indicate unacceptable, poor, average, and good differentiation indexs, respectively. These quantitative indicators of classical test theory (difficulty index and discrimination index) were obtained usage of the Excel forms in the test center for each test.

The study was approved by the Ethics Committee of the Research Council of Hamadan University of Medical Sciences under the number IR.UMSHA.REC.1398.740. Since no human samples were used and the data were extracted from the forms available in the test center, no individual consent was required. Data were analyzed using SPSS software version 20. Mean and standard deviation was used to report quantitative variables according to the normal distribution of data. Ratios and percentages were also reported to report qualitative variables.

Results

All 200 questions of the last three exams in 2018, 2019 & 2020 were examined. These questions were designed in 17 different subjects (by 17 educational groups). The number of participants in 2018, 2019 & 2020 in Hamadan University of Medical Sciences were 24, 35 and 86, respectively. The average validity of the tests in 2018-2020 was 0.86, 0.91, and 0.88, respectively. All three tests had good validity. The average discrimination index in 2018-2020 was equal to 0.19, 0.09, and 0.21, respectively, which shows that the discrimination index was moderate in 2018 and 2020 and weak in 2019 (Table 1).

[Downloaded from edujournal.zums.ac.ir on 2026-11-09]

[DOI: 10.66224/edcj.14.44.36]

Table 1: Frequency distribution of questions with high taxonomy, without structural defects, difficulty and discrimination index of questions in medical pre-internship exams for the years 2018 to 2020

Variables	September 2018 (Percentage) number	September 2019 (Percentage) number	September 2020 (Percentage) number
Questions with high taxonomy	116(58%)	127(63.5%)	127(63.5%)
Questions without structural problems	182(91%)	181(90.5%)	›96%‹192
Number of questions with negative discrimination index	41(20.5%)	39(19.5%)	15(7.5%)
Number of questions with poor discrimination index	24(12%)	84(42%)	37(38.5%)
Number of questions with medium discrimination index	100(50%)	75(37.5%)	133(66.5%)
Number of questions with good discrimination index	35(17.5%)	2(1%)	15(7.5%)
Number of easy questions	55(27.5%)	46(23%)	35(17.5%)
Number of appropriate questions	107(53.5%)	112(56%)	122(61%)
Number of hard questions	38(19%)	42(21%)	43(21.5%)

The average difficulty index in 2018, 2019 and 2020 were 0.53, 0.51, and 0.49, respectively, indicating that the questions in each of the three years were at appropriate level in terms of difficulty (Figure 1). The number of easy questions in medical pre-internship exams in 2018, 2019 and 2020 was 55, 46 and 35, respectively and the number of appropriate questions was increased (Table 1).

More than 90% of the questions in these years were without structural defects and in 2020 the least structural defects were observed in the questions. In 2018, 116 questions (56% of questions) and in 2019 and 2020, 127 questions (64% of questions) were designed at a high taxonomic level (2 and 3) (Table 1).

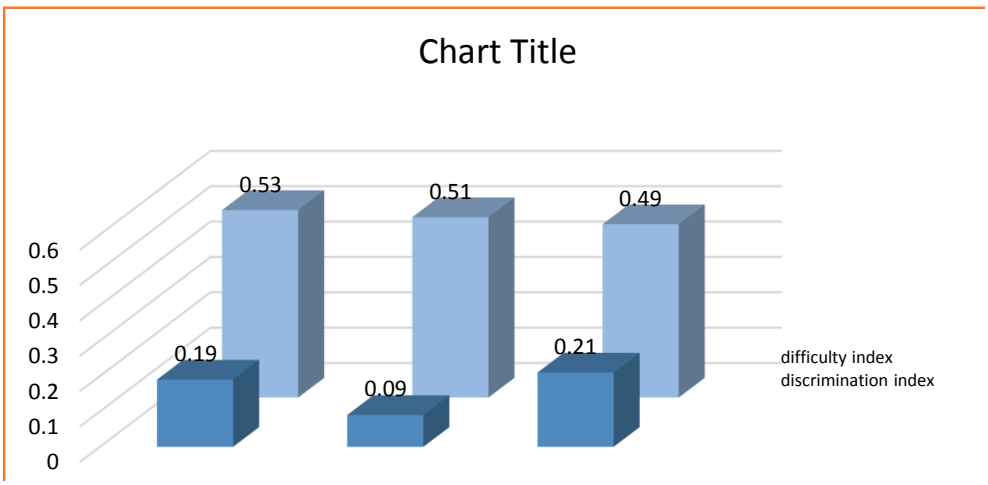


Figure 1: Distribution of the average difficulty index and discrimination index of medical pre-internship exam questions for the years 2018 to 2020

[Downloaded from edujournal.zums.ac.ir on 2026-11-09]

[DOI: 10.66224/edcj.14.44.36]

Discussion

The general results of this study showed that the level of difficulty of the tests in the years under study was moderate (about 0.50). The average discrimination index of questions in the studied years was moderate and weak (between 0.09 and 0.21). A closer look at the results shows that more than 60% of the questions in 2019 had a negative or weak differentiation index, and in 2018 and 2020 this value was about 32% and 25%. Since the closer the differentiation index of each test is to one, the test has better discrimination index and can differentiate between strong and weak students in the course, so the tests in question, especially in 2019, are well able to differentiate. Weak students have not been strong students. For the clean factor to improve, the difficulty level of the questions must be desirable because difficult and very difficult questions, as well as very easy questions, have little resolution (8). If the level of difficulty of the questions is desirable, the pattern of answers and the text of the questions, and deviant options should be considered. Questions with a negative discrimination index indicate that in that question, the weak group performed better than the strong group.

Such questions have fundamental flaws that need to be removed or fundamentally revised.

There is a lot of ambiguity in the body of the question or options.

Other reasons for the possibility that the key option announced by the question designer is incorrect or that the correct answer key is entered incorrectly in the answer sheet (8, 3). A look at different studies in designing questions for similar tests reveals similarities and differences. Nibret et al. In their study on the analysis of 176 first-year students to the same results as the present study with a discrimination index of 0.16 and a difficulty index of 0.56. (15) In a study conducted by Pourmirza Kalhori et al. In 2013 on the five-year process of designing medical residency promotion exams for Kermanshah University of Medical Sciences, the total difficulty index of the questions was estimated at 62% and the question differentiation index at 27%, both of which were in the desired range (10) In the Anbari study, in the residency promotion exams of 2012, 44.7% of the questions were evaluated in terms of difficulty index, and 54.1% of the questions were evaluated in terms of appropriate differentiation index (16). In the study by Khafagy et al In Egypt, the clean factor of questions increased in 2013 compared to 2009 and in 2013 it was higher than 0.3, and the difficulty of questions was reduced from 65 to 55 (17). Since the content of the questions are different, different results are also expected, however, it seems that comparing the psychometric process of the tests over time is more important. The results of this study also show that about 58% of the questions were designed in 2018 with a high-level taxonomy, which has reached 64% in subsequent years. Also in all three tests, the design of most of the questions was structurally flawless. Comparison of these results with similar internal or external studies shows that these tests, which are related to the beginning of students entering the internship stage and students' responsibility to patients, have been designed at an acceptable level in terms of taxonomy. The study of Saburi et al. In 2019 in a 5-year study of the questions of the specialized board of disciplines related to adult cancer from 2013 to 2017 showed that 54% of the questions were designed at taxonomy level 1 and 23% at the level of taxonomy 2(18) In the study of Pourmirza Kalhori et al., On average, 33.4% of the questions were with taxonomy level 1 and 66.6% were questions with taxonomy level 2 and 3 and 62.6% of the questions were designed without structural defects (10). Shakornia's study of residency promotion test questions in Jundishapur in 2010 showed that only 7% of the questions were designed at taxonomy level 2 and 3 (19). The study of Anbari et al. The evaluation of the promotion test for residents of clinical departments of Arak University of Medical Sciences in

[Downloaded from edujournal.zums.ac.ir on 2026-11-09]

[DOI: 10.66224/edcj.14.44.36]

2012 showed that 80% of the questions were designed without structural defects and about 65.2% of the questions were designed at taxonomy level 2 and 3 (16). In the study of Baqaei et al., that examined the level of observance of multiple-choice questions in the Faculty in Nursing and Midwifery of Urmia University in 2015. The results showed that 85% of the questions were at the level of a Bloom taxonomy and in general, the questions of Bloom taxonomy weren't at the desired level. (20) Comparison of the results of this study with other studies shows that the design of questions in terms of structural defects and taxonomy in questions is designed at a good and acceptable level, however, it seems that question designers are better in addition to paying attention to taxonomy and correct structure in questions and The level of difficulty of the tests, in terms of selecting the appropriate deviation options, have passed the necessary training, and in addition, after designing the test, the questions should be reviewed preferably by expert and trained colleagues to select the appropriate question and deviation options and review the key questions.

As standard and in their study in 2012 showed that about 35% of residency promotion questions had structural problems that with the intervention this amount was significantly reduced and also the number of questions with high taxonomy was increased from 38% to 53% (21). The results of this study can be used to improve or enhance the quantity and quality of pre-internship questions for medical students of medical universities. Teachers' knowledge in evaluating and evaluating is a dynamic and continuous activity. So, it is recommended that professors or instructors participate in workshops around test developments and test psychometrics and update their knowledge. Also, to minimize the questions with a negative differentiation index before the test, the questions should be reviewed by the group colleagues to correct their problems. Since most of the exams made in the faculties and by the professors are of the objective type, the analysis of the exams or at least some of them is recommended for the teachers. It is also suggested that due to the existence of an electronic test system and the possibility of extracting test analysis, a special kind of test management should be considered, that is responsible for analyzing questions and responsible for feedback to professors and even guidance and advice on improving design and taxonomy and test questions.

Conclusion

The results of this study showed that the pre-internship test questions from 2018 to 2020 had a better quality in terms of quality indicators than previous years. As the questions are designed with a higher taxonomy and the designers of the question are required to observe the correct principles of question structure. However, in terms of quantitative indicators, the existence of a large number of questions with a weak and negative differentiation index indicates the need for re-examination. Therefore, it is recommended to establish training sessions and provide appropriate and appropriate feedback before designing the questions, and after designing the questions, deviant options and correct answers should be reviewed again.

Acknowledgments

This study has been carried out with the approval of the Vice Chancellor for Research and Technology of Hamadan University of Medical Sciences with design number 9809267144 and the Ethics Committee of the Research Council of Hamadan University of Medical Sciences under the number IR.UMSHA.REC.1398.740.The authors of this article would like to express their gratitude and appreciation to the experts of the Examination Center of Hamadan University of Medical Sciences who, while observing the confidentiality of the test results, expressed their utmost cooperation in providing the required test questions and indicators.

References

1. Iqbal MZ, Khan RA, Razaq N. Assessment of non-functional distracters in multiple choice questions: a descriptive analysis. *Pak J Physiol.* 2016;12(2):47-9.

[Downloaded from edujournal.zums.ac.ir on 2026-11-09]

[DOI: 10.66224/edcj.14.44.36]

2. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *JPMA-J Pak Med. Assoc.* 2012;62(2):142-147.

3. Saif A. Educational measurement, assessment and evaluation. Tehran: Doran Publications. 2004;128.

4. Amin Z, Chong YS, Khoo HE. Practical guide to medical student assessment:World Scientific; 2006.

5. Roediger III HL, Marsh EJ. The positive and negative consequences of multiple-choice testing. *J Exp. Psychol: Learn. Mem. Cognit.* 2005;31(5):1155-59.

6. Talebi GA, Ghaffari R, Eskandarzadeh E, Oskouei AE.. Item Analysis an Effective Tool for Assessing Exam Quality, Designing Appropriate Exam and Determining Weakness in Teaching. *Res Dev Med Educ.* 2013; 2(2): 69-72.

7. Abbasi H, Falsafinejad MR, Delavar A, Farrokhi NA, Mohagheghi MA. The Comparison of Two Models for Evaluation of Pre-internship Comprehensive Test: Classical and Latent Trait. *Iran. J Med. Educ.* 2013;13(3):167-78.

8. Amin MM, Shayan S, Hashemi H, Poursafa P, Ebrahimi A. Analysis of multiple choice questions based on classical test theory. *Iran. J Med. Educ.* 2011;10(5):719-25(persian).

9. pourmirza kalhori R, rezeai M, Karami Matin B, Roshan Pour F. A survey of quality and quantity indexes of multiple choice question (MCQ) exams of medical residents at Kermanshah University of Medical Sciences: 2008-2012. *J Med. Educ. Develop.* 2014; 8 (4) :64-75

10. Haladyna T M, Downing S M, Rodrigues M C. A review of multiple choise item writing guidelines for classroom assessment. *Appl. Measurement Educ.* 2002; 15(3):309-334.

11. Zaidi NLB, Grob KL, Yang J, Santen SA, Monrad SU, Miller JM, et al. Theory, process, and validation evidence for a staff-driven medical education exam quality improvement process. *Med Sci Educ.* 2016;26(3):331-6.

12. Khazaei MR, Zarin A, Rezaei M, Khazaei M. Factors affecting the results of comprehensive pre-internship exam among medical students of Kermanshah University of Medical Sciences. *Korean j med. educ.* 2018;30(2):131-139.

13. Musa A, Shaheen S, Ahmed A. Distractor analysis of multiple choice questions: A descriptive study of physiology examinations at the Faculty of Medicine, University of Khartoum. *Khartoum Med. J.* 2018;11(1).

14. Jafari F, Hadavand M, Samadpour M, Azami FH, Behbahan SEB. Evaluation of Pre-Internship Comprehensiv Exams Scores and their Predictive Factors. *Biomed Pharmac J.* 2013;6(2):307-13.

15. Nibret TK. Post exam analysis: Implication for intervention. *bioRxiv.* 2019; 1:510081.

16. Anbari Z, Jadidi R. Assessment of the resident's promotion exam: One step to validity of competency measurement in Arak University of Medical Sciences.*J med. educ. develop.* 2013; 7 (4) :52-62

17. Khafagy G, Ahmed M, Saad N. Stepping up of MCQs' quality through a multi-stage reviewing process. *Educ Primary Care.* 2016;27(4):299-303.

18. Sabouri M, Arbabi F, Dehghani Poudeh M. The Investigation into the Board Examinations in Majors Related to Adult Cancer in Iran. *Iran. J Med. Educ.* 2020; 20 :272-278.

19. Shakurnia A, Mozaffari A, Khosravi Brougeni A. Survey on Structural of MCQs of residency exam in AJUMS. *Jundishapur Sci Med J.* 2010;8(4):492-502.

20. Baghaei R, Feizi A, Shams S, Naderi J, Rasouli D. Evaluation of the nursing students' final exam multiple-choice questions in Urmia University of medical sciences. *J Urmia Nurse. Midwifery Fac.* 2016;14(4291):291-9.

21. Meyari A, Beiglarkhani M, Zandi M, Vahedi M, Miresmaeili A. The Effect of Education on Improvement of Multiple Choice Questions' Designing in Annual Residency Exams of Dental School. . *Iran. J Med. Educ.* 2012;12(1):36-45.

Rostami-Moez M, Biglarkhani M, Meyari A. Quality of design and analysis of pre-internship exams for medical students of Hamadan University of Medical Sciences in the years 2018 to 2020. J Med Educ Dev. 2022; 14 (44):36-42

[Downloaded from edujournal.zums.ac.ir on 2026-11-09]

[DOI: 10.66224/edcj.14.44.36]