Evaluation of Anatomy Multiple Choice Questions for First and Second-year Students in the College of Medicine, University of Mosul

Rana M. Raoof*
***Department of Anatomy, College of Medicine, University of Mosul, Mosul , Iraq**
Correspondence: rmr@uomosul.edu.iq

(Ann Coll Med Mosul 2024; 46 (1):128-136).
Received: 6th Febr. 2024; Accepted: 3rd April 2024.

ABSTRACT

Background: The ability to generate a high-quality multiple-choice question (MCQ) is a very important skill for every medical educator. Analysis of these questions is an important post-test evaluation step that will give valuable feedback to the item (MCQ) constructor regarding how difficult and discriminative each item was. Moreover, the effectiveness of each alternative is also investigated and their efficiency is calculated.

Objectives: To evaluate the quality of MCQs used in Anatomy exam over a period of 2 years.

Materials and methods: A cross-sectional study was conducted to analyze 320 MCQs used in four anatomy tests. For each item, the difficulty index (DIF), discrimination index (DI) and distractor efficiency (DE) were calculated.

Results: The mean difficulty of the four tests ranges between 57.35-61.52. The majority of MCQs were either of average difficulty (39.7%) or moderately easy (35.3%). Seventy per cent of MCQs were highly discriminative with DI above 0.4. The efficiency of the distractors was 100% in 49.7% of questions. The four tests were highly reliable with KR-20 >0.9. A very strong negative and positive correlation was found between DIF and DE, DI and DE respectively.

Conclusion: The four anatomy tests showed high reliability and acceptable difficulty and discrimination reflecting competence in item writing. However, Conscious attention is always required while writing the distractors to eliminate any non-functioning ones thus Increasing the DE of the MCQ.

Keywords: Anatomy, item analysis, difficulty, discrimination, distractor efficiency.

تقييم الاسئلة متعددة الاختيارات لمادة التشريح لطلبة السنة الأولى والثانية في كلية الطب جامعة الموصل

رنا ممتاز رؤوف*
*فرع التشريح ، كلية الطب ، جامعة الموصل ، الموصل ، العراق

الخلاصة

الخلفية: تعد القدرة على إنشاء سؤال متعدد الاختيارات عالي الجودة مهارة مهمة جدًافي مجال التعليم الطبي. يعد تحليل هذه الأسئلة خطوة مهمة في تقييم ما بعد الاختبار والتي ستوفر معلومات قيمة فيما يتعلق بمدى صعوبة كل عنصر وتمييزه. علاوة على ذلك، يتم أيضًا دراسة فعالية كل مشتت وحساب كفاءته.

الأهداف: تقييم جودة الأسئلة MCQs المستخدمة في امتحان التشريح على مدى عامين.

المواد والطرق: أجريت دراسة مقطعية لتحليل ٣٢٠ سؤالاً متعدد الأختيارات. لكل عنصر، تم حساب مؤشر الصعوبة (DIF)، ومؤشر التمييز (DI) وكفاءة التشتيت (DE).

النتائج: متوسط صعوبة الاختبارات الأربعة يتراوح بين ٥٧.٣٥-٦١.٥٢. وكانت غالبية الأسئلة MCQ إما متوسطة الصعوبة (٣٩.٧%) أو سهلة إلى حد ما (٣٥.٣%). ٧٠.٣% من الأسئلة MCQs كانت تمييزية للغاية مع DI أعلى من ٠.٤. وكانت كفاءة المُشتتات ١٠٠% في ٤٩.٧% من الأسئلة. وكانت الاختبارات الأربعة موثوقة مع KR-20> ٠.٩. تم العثور على علاقة سلبية وإيجابية قوية جداً بين DIF وDE، DI وDE على التوالي.

الاستنتاج: أظهرت اختبارات التشريح الأربعة موثوقية عالية وصعوبة مقبولة وتمييزاً يعكس الكفاءة في كتابة المادة. ومع ذلك، مطلوب دائمًا الاهتمام الواعي أثناء كتابة عناصر التشتيت لإزالة أي عناصر غير عاملة وبالتالي زيادة DE الخاص بـ MCQ.

الكلمات المفتاحية : التشريح، تحليل العناصر، الصعوبة، التمييز، كفاءة التشتيت.

INTRODUCTION

Assessment of students' knowledge and performance is a cornerstone process that determines progress in medical schools. The quality assurance of all the assessment practices from planning, preparation, execution, and analysis of examination results is thus essential to ensure valid and reliable assessment practices [1] . Multiple Choice Questions (MCQs) are commonly used to test knowledge outcomes in medical colleges due to their preference among both students and faculty members [2,3] . When appropriately constructed, MCQs (commonly referred to as items) are superior to essay questions as they can test the student's higher cognitive abilities especially, the application, interpretation and synthesis of knowledge [4] .

Constructing an effective MCQ test needs a pre-test conscious decision from the examiner regarding the test difficulty and test discrimination abilities of each item [5] . Still, post-test psychometric evaluation of the MCQ items can give valuable information about the quality of the test. This is performed mainly by calculating the DIF, the DI and DE for each item used in the test, a process known as item analysis [6] .

The item difficulty index (DIF) reflects the number of students who correctly answered the item and it ranges between 0-100%. A smaller DIF, means a more difficult question [7] . On the other hand, the item discrimination index (DI) is defined as the ability of the item to distinguish between high and low-achieving students within the same cohort. It ranges between -1 and +1. Ideally, a perfect item will have a DI of +1, which means that all the high-achieving students and none of the low-achieving students have chosen the correct answer [7] . However, this is not the case in the real world and the DI of a properly constructed item usually ranges between 0 - >0.4 [6,8-10]

Another item analysis parameter is the distractor efficiency (DE). Each item contains a stem and four to five options, one is the correct option (key). All the other options are incorrect alternatives and are called distractors. Non-functional distractors (NFD) are those alternatives selected by less than 5% of students. Every distractor selected by 5% or more students is regarded as a functional distractor (FD) [11] . Based on how many FDs each item has, its DE will range from 0% to 100%. If all the distractors were efficient (functional) then the DE is 100%. If one, two, three or four distractors were regarded as NFD then the DE would be 75%, 50%, 25% and 0% respectively [6,12] .

Test reliability is another quality assurance parameter. It measures the internal consistency of the test results. In other words, how well the test is measuring what we want to measure. Many formulas have been used to calculate the internal consistency (Reliability) of a test for example Cronbach alpha and Kuder Richardson (KR)-20 [13] . The test reliability score ranges between 0 to 1. The closer the number to 1 means better reliability and at least a score of 0.7 is required for an MCQ test to be considered reliable [14] .

The study aimed to analyze the quality of the Anatomy MCQ questions used in Anatomy final exams over two academic years (2021-2022 and 2022-2023).

MATERIAL AND METHODS

Data Collection and Ethical Approval:

The study was conducted in the Department of Anatomy and was approved by the College Council who authorized access to the examination data. Students' identities were kept anonymous and confidential at all stages of the study.

A cross-sectional descriptive study was designed to analyze 320 Anatomy MCQ items with their 1280 distractors. These items were used for the final exams for the Anatomy I course and the Anatomy II course that were conducted over the years 2022 and 2023. The characteristics of each test (number of students, number of items, mean and range of test scores) are shown in Table 1. In the four tests, each item consisted of a stem and five different answers from which the students were asked to select the correct one. Student exam papers were automatically corrected using an optical Mark Recognition system.

Table 1: Number of students, number of items, mean (± SD) and the range of test scores in each final exam for the years 2021-2022 and 2022-2023

Test	Academic year	No. of students	Course	No. of items	Mean test score ± SD	Range of test score
1	2021-22	427	Anatomy I	100	61.52 ± 19.3	16-100
2	2021-22	690	Anatomy II	100	57.35 ± 16.7	9-98
3	2022-23	497	Anatomy I	60	35.55 ± 10.9	8-57
4	2022-23	452	Anatomy II	60	35.22 ± 11.03	9-59

Item Analysis and Data Interpretation

Students were first ranked based on their test scores; each correct answer gets 1 point. Data obtained were then analyzed using Microsoft Office Excel and GraphPad Prism software. Statistical significance is determined using ANOVA, with $P<0.05$ considered to be statistically significant.

The DIF: was calculated for each item using the formula (DIF=C*100/TN, where C=total correct responses for that particular item and TN=total number of students) [11] . Interpretation of DIF according to the Local college guidelines is shown in Table 2A.

The DI: was calculated for each item using the formula (DI=HG-LG/n). where HG= The number of students who answered the item correctly in the high-achieving group (top 25%) and LG= The number of students who answered the item correctly in the low-achieving (LG) group (lower 25%), and n=total number of students in each group [11] . Interpretation of DI according to the Local college guidelines is shown in Table 2B.

Table 2: Interpretation of **(A)** DIF and **(B)** DI according to local guidelines

DIF %	
80.01-100	Very easy
60.01-80	Moderately easy
40.01-60	Average
20.01-40	Moderately difficult
0-20	Very difficult

DI	
<0	Faulty question, eliminate
0-0.19	Poor discrimination
0.20-0.29	Acceptable discrimination
0.30-0.40	Good discrimination
>0.40	Excellent question

DIF; Difficulty index, DI; Discrimination index.

Test reliability: was calculated based on Kuder-Richardson (KR)-20 formula as it is specifically used for items with binary data (0,1) [15] using the formula:

KR-20= (k / (k-1)) * (1 – Σpiqi / σ2). where: k= Number of MCQs in the test, pi= Proportion of students who correctly answered MCQ (j), qi= Proportion of students who incorrectly answered MCQ (j), σ2= score variance for all students who took the test.

The number of NFD and DE was calculated for each item [5] . A distractor is considered a NFD if it was selected by fewer than 5% of students. The correlation between different item analysis parameters was calculated using Pearson's correlation coefficient (r). The correlation is considered very weak, weak, moderate, strong or very strong based on r values (between 0-0.19, between 0.2–0.39, between 0.40–0.59, between 0.6–0.79 and between 0.8–1 respectively) [16] .

RESULTS

The four tests were highly reliable with KR- 20 above 0.9 (Table 3). A total number of 320 Items, were analyzed with the mean DIF for the 320 items being 59.26±16.7. The hardest item in the four tests had a DIF of 16.4% (which means only 16.4% of students were able to choose the correct answer) while the easiest item had a DIF of 97.2%. The number of questions in each difficulty category is shown in Table 3.

Table 3: Reliability, mean DIF and No. (%) of items in each category based on DIF interpretation according to local guidelines.

Test	Test reliability (KR-20)	Mean DIF ± SD	No. (%) of items				
			VE	ME	A	MD	VD
1	0.95	61.52 ± 15.7	10 (10.0)	40 (40.0)	39 (39.0)	11 (11.0)	0 (0.0)
2	0.93	57.35 ± 17.2	10 (10.0)	32 (32.0)	40 (40.0)	16 (16.0)	2 (2.0)
3	0.91	59.25 ± 17.1	7 (11.7)	22 (36.7)	22 (36.7)	8 (13.3)	1 (1.7)
4	0.90	58.69 ± 17.2	8 (13.3)	19 (31.7)	26 (43.3)	6 (10.0)	1 (1.7)
Total			35 (10.9)	113 (35.3)	127 (39.7)	41 (12.8)	4 (1.3)

A; Average, DIF; Difficulty index, ME; moderately easy, MD; moderately difficult, VE, very easy, VD; very difficult.

Of the 320 test items analyzed, the majority of questions were found to be either of average difficulty (127 MCQs, 39.7%) or moderately easy (113 MCQs, 35.3%). The number of very easy questions (35 MCQs, 10.9%) was near to that of the moderately difficult ones (41 MCQs, 12.8%). Only a few MCQs were considered very difficult (4 MCQs, 1.3%) (Figure 1A). The 4 tests were comparable in regard to their difficulty and no significant difference was found (Figure 1B).

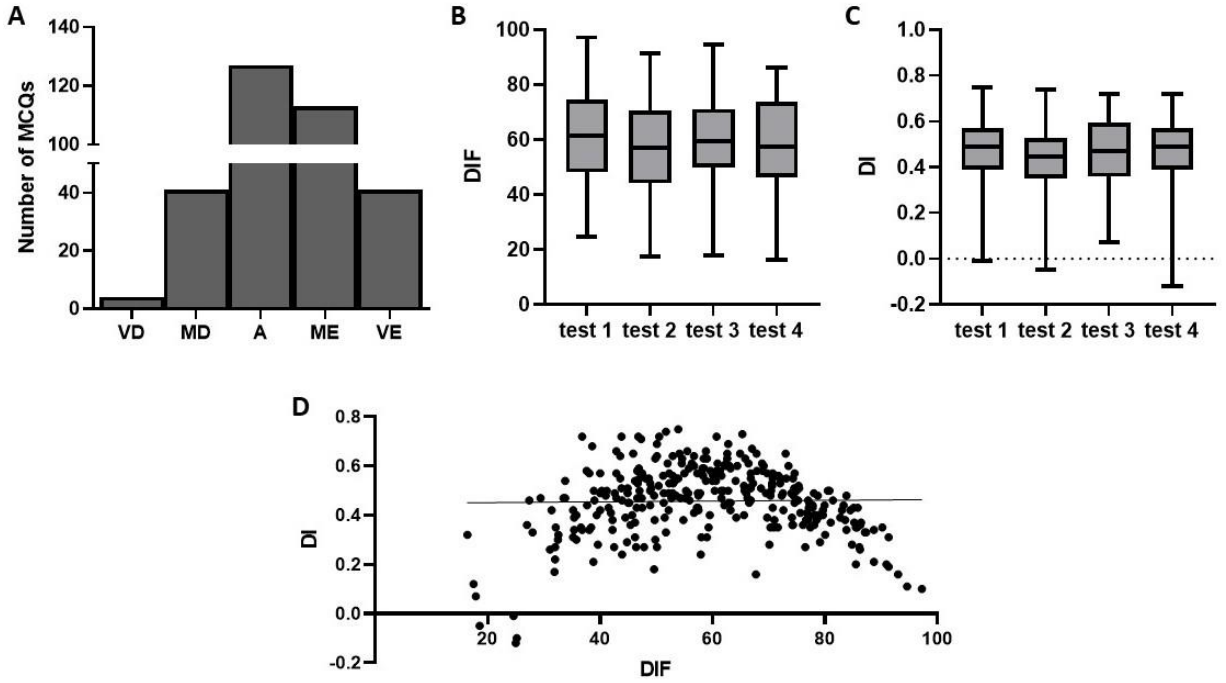

Figure 1: (A) Frequency distribution of the number of MCQs according to their difficulty levels for the four tests. (B) box plot showing the mean DIF for the 4 tests. (C) box plot showing the mean DI for the 4 tests. (D) Correlation between the DIF and the DI for the 320 MCQs analyzed. A; Average, DIF; Difficulty index, DI; Discrimination index, ME; Moderately Easy, MD, Moderately Difficult, VE; Very Easy, VD; Very Difficult

Of the total 320 items analyzed, the majority (over 70%) have an excellent DI. Only a few were regarded as faulty or having poor discrimination (1.3% and 2.8% respectively) (Table 4). No significant difference was found between the discriminative abilities of the 4 tests (figure 1C). Analysis of the correlation between DIF and DI of the 320 items showed a very weak positive correlation (r=0.0184) of these 2 parameters (Figure 1D). The correlation analysis using individual data from each was also performed and this also showed a very weak negative correlation that is not significant (data not shown).

Table 4: Mean ± SD of DI and No. (%) of items in each category based on DI interpretation for the four tests according to local college guidelines.

Test	Mean DI ± SD	No. (%) of items				
		Faulty question	Poor DI	Acceptable DI	Good DI	Excellent DI
1	0.48 ± 0.14	1 (1.0)	3 (3.0)	5 (5.0)	18 (18.0)	73 (73.0)
2	0.43 ± 0.13	1 (1.0)	3 (3.0)	9 (9.0)	24 (24.0)	63 (63.0)
3	0.46 ± 0.15	0 (0.0)	3 (5.0)	4 (6.7)	7 (11.7)	46 (76.7)
4	0.47 ± 0.15	2 (3.3)	0 (0.0)	2 (3.3)	13 (21.7)	43 (71.7)
Total		4 (1.3)	9 (2.8)	20 (6.2)	62 (19.4)	225 (70.3)

DI; Discrimination index, SD; Standard deviation.

Analyzing the effectiveness of 1280 distractors for the 320 items was also performed. The number and frequency of Functional distractors (FD) were reported (Table 5). Nearly three-quarters of the items had an excellent distractor efficiency of either 100% or 75% reflecting very high-quality questions. However, items with 2,1 and 0 functional distractors represented 12.5%, 9.4% and 3.4% respectively. Frequency distribution analysis of the number of items according to their DE was done for each test individually (Figure 2). Nearly half of the questions in the four tests have 100% DE.

Table 5: Number of items in each distractor efficiency category with the number of their functional and non-functional distractors of the four Anatomy exams.

DE Categories	Number of items (%)
items with 4 functional distractors (0 NFD, DE =100%)	159 (49.7)
items with 3 functional distractors (1 NFD, DE =75%)	80 (25.0)
items with 2 functional distractors (2 NFD, DE =50%)	40 (12.5)
items with 1 functional distractor (3 NFD, DE =25%)	30 (9.4)
items with 0 functional distractors (4 NFD, DE =0%)	11 (3.4)

DE; Distractor efficiency, NFD; non-functioning distractor

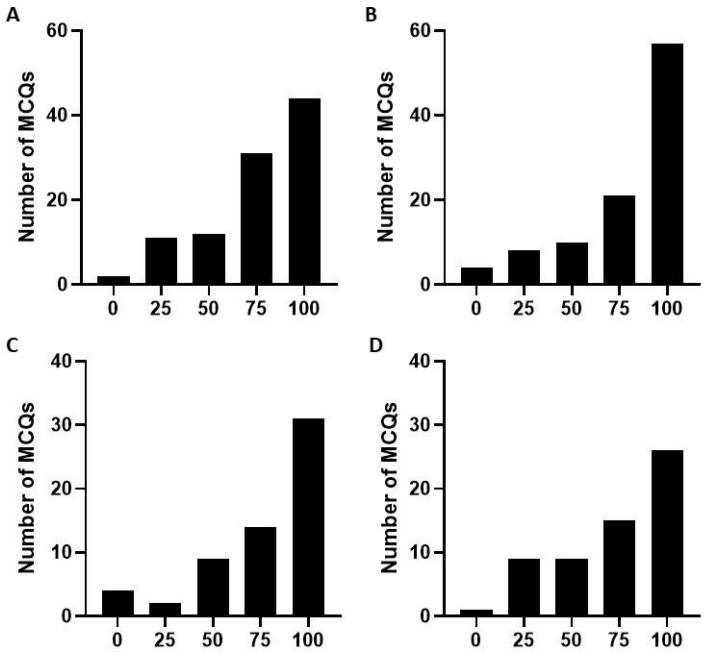

Figure 2: Frequency distribution of the MCQs according to their DE for test 1(A), test 2 (B), test 3 (C) and test 4 (D). DE; Distractor Efficiency

The correlation between the DE of the 320 items and their DIF and DI was then analyzed. DE was significantly negatively and positively correlated with the DIF (Figure 4E) and DI (Figure 5E) respectively. This means that the more efficient distractors the item has, the more difficult and more discriminative it will be. The analysis was also repeated on each test. While the significant negative correlation between DIF and DE was maintained between the 4 tests (Figure 3, A,B,C,D), the significant positive correlation between the DI and DE was only maintained in tests 1,2, 3 and it was lost in tests 4 (Figure 5, A,B,C,D).

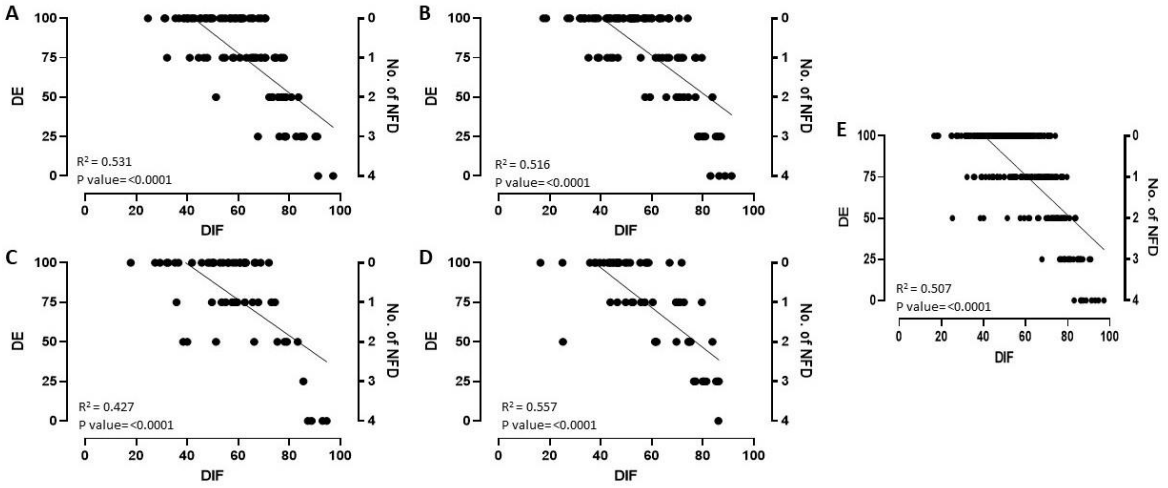

Figure 3: Correlation between distractor efficiency (DE) and the number of non-functioning distractors (NFD) in the Y-axis, and difficulty index (DIF) in the X-axis, for each MCQ item in test 1 (A), test 2 (B), test 3 (C) and test 4 (D). (E) is the correlation between the DE and DIF of the 320 MCQ items analysed in the four tests.

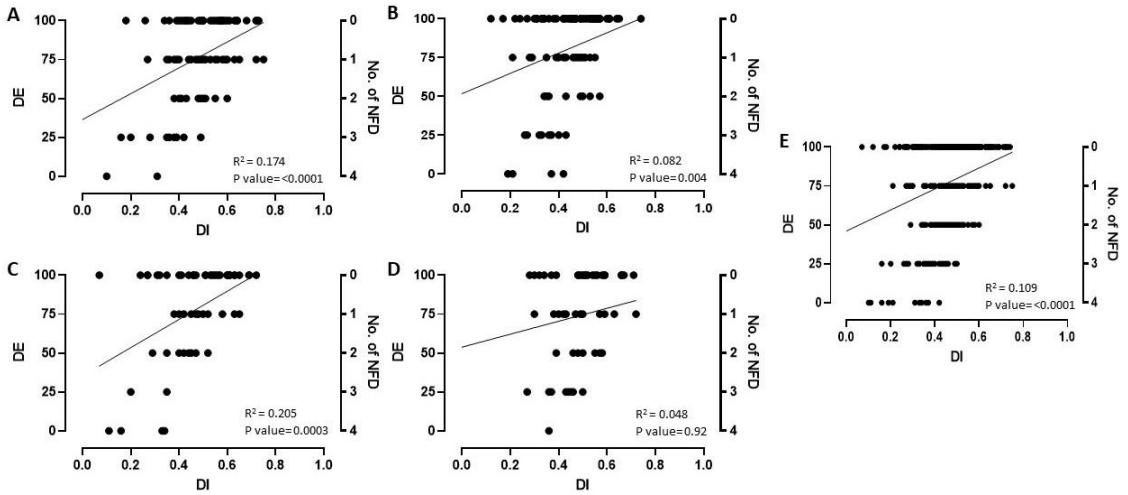

Figure 4: Correlation between distractor efficiency (DE) and the number of non-functioning distractors (NFD) in the Y-axis, and the discrimination index (DI) in the X-axis, for each MCQ item in test 1 (A), test 2 (B), test 3 (C) and test 4 (D). (E) is the correlation between the DE and DI of the 320 MCQ items analysed in the four tests.

DISCUSSION

Item analysis is the process by which the performance of different items included in the exam is assessed based on the student's response. This analysis will help in evaluating the quality of each item as well as the quality of the exam as a whole, thus, increasing confidence in test scores [17] . Item analysis will also give feedback to teachers about the effectiveness of their teaching practices, the need to give more focus on certain outcomes that students found difficult and the teacher's item writing skills. In addition, the different item statistics measured are important not only to decide the reliability of the exam but also to help to decide which items are to be retained in the item bank (good items) and which items need further improvement of the stem and/or distractors (badly performing items) [11] .

Ideally, all MCQ questions used in any summative test should have been evaluated previously, formatively, in order to determine each MCQ difficulty and discrimination indices. This allows the examiner to use the appropriate difficulty level for the whole test and to use the items that best discriminate between high and low-achieving students. A perfectly constructed item should have a DIF between 30–70%, DI of >0.25 and a DE of 100% [18].

The majority of the items analyzed in this study were of an acceptable DIF (DIF is between 30-70%). The mean DIF of 4 analyzed tests is considered acceptable by many other studies which reported mean exam difficulty of 53.2±22.4 [19], 54.1±17.5 [20], 47.9±16.4 [21]. In their paper, Kheyami et al [18] analyzed 16 summative exams performed over 3 years (2013-2016). The mean DIF of these exams ranged between 36.70–73.14% which they considered acceptable. They reflect the considerable variation in item DIF to continuous improvement in faculty item writing skills. Another study by Karelia et al [22] analyzed 12 summative exams administered between 2008-2012, they reported a mean DIF of the individual tests in the range of 47.17% to 58.08%.

The four tests had a very high ability to discriminate between high and low-achieving students. The mean DI of the analyzed tests ranged between 0.43-0.48. More than 70% of the 320 questions were of excellent DI (>0.4). This reflects a very good ability among faculty members in regard to item construction and a high confidence that these tests can differentiate between good and bad students. This is much higher than the discrimination abilities of other tests reported in literature where studies have reported only 0% [21], 29% [23], 32.5% [19] of items have an excellent DI. However, other studies had a comparable level of items with excellent discrimination to our study [17,20]. Items with poor or negative DI usually have incorrect answer keys or are suffering from unclear, confusing stems. These items should not be utilized again unless they are corrected [18].

Despite one might think that there should be a strong correlation between the item DIF and DI; where more difficult questions can more effectively discriminate between low-achieving and high-achieving students and vice versa; the present work showed a very weak non-significant correlation which was also reported by other studies [19,21]. This weak negative correlation was reported to be significant by Mitra et al [24]. Other studies also reported a weak, but significant positive correlation between DIF and DI with the maximal discrimination occurring with moderately easy to moderately difficult items (DIF between 40-70%) [9,18,23].

Having plausible distractors is another measure of good MCQ. An item with DE of 0,25 or 50% reflects either a poorly constructed item or an easy item that is utilized by the examiner on purpose. In the current study, nearly half of the items had a 100% DE. Other studies reported 13.3%,13.8% and 70% [5,21,25]. In the current study, the number of items with 100% DE was highly negatively correlated with the DIF and positively correlated with DI. This means that the higher the number of functional, plausible distractors the item has, the more difficult and discriminative it will be. Any increase in the number of nonfunctional distractors will reduce the DI and increase the DIF of the item so the item will be easier to be answered by a larger number of students.

CONCLUSION

Performing an item analysis is an important step after every MCQ exam. It will ensure a high-quality, reliable exam that can effectively differentiate between low and high achieving students. Item analysis parameters also very important to develop an MCQ bank which contain items with acceptable DIF and DI. Feedback to teachers should also be given after the test on how their MCQs performed and any possible improvement measures.

Source(s) of Support

The author did not receive any type of support to complete the paper

Acknowledgement

N/A

Ethical Approval

The CMUM Council approved the study and authorized access to the examination data. Students' identities were kept anonymous and confidential at all stages of the study.

Conflicts of Interest

The author has no conflicts of interest to declare.

REFERENCES

1. Norcini JJ, McKinley DW. Assessment methods in medical education. Teach Educ. 2007 Apr 1;23(3):239–50. Available from: https://doi.org/10.1016/j.tate.2006.12.021.
2. Neto J, Neto F, Furnham A. Predictors of students' preferences for assessment methods. Assess Eval High Educ. 2023;48(4). Available from: https://doi.org/10.1080/02602938.2022.2087860

3. Furnham A, Batey M, Martin N. How would you like to be evaluated? The correlates of students' preferences for assessment methods. Pers Individ Dif. 2011;50(2). Available from: https://doi.org/10.1016/j.paid.2010.09.040.
4. Peitzman SJ, Nieman LZ, Gracely EJ. Comparison of "fact-recall" with "higher-order" questions in multiple-choice examinations as predictors of clinical performance of medical students. Academic Medicine. 1990;65(9). Available from: http://journals.lww.com/00001888-199009000-00044
5. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. BMC Med Educ. 2009;9(1). Available from: https://bmcmededuc.biomedcentral.com/articles/10.1186/1472-6920-9-40
6. Kumar D, Jaipurkar R, Shekhar A, Sikri G, Srinivas V. Item analysis of multiple choice questions: A quality assurance test for an assessment tool. Med J Armed Forces India [Internet]. 2021 Feb 1 [cited 2023 Jul 14];77(Suppl 1):S85–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/33612937
7. McCrossan P, Nicholson A, McCallion N. Minimum accepted competency examination: test item analysis. BMC Med Educ. 2022;22(1). Available from: https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-022-03475-8.
8. Sarin YK, Khurana M, Natu M V., Thomas AG, Singh T. Item Analysis of Published MCQs. Indian Pediatr. 1998;35(11). Available from: https://pubmed.ncbi.nlm.nih.gov/10216545/
9. Sim SM, Rasiah RI. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med Singap. 2006;35(2).). Available from: https://annals.edu.sg/pdf/35VolNo2200603/V35N2p67.pdf
10. Patil R, Palve S, Vell K, Boratne A. Evaluation of multiple choice questions by item analysis in a medical college at Pondicherry, India. Int J Community Med Public Health. 2017; 3(6), 1612–1616. Available from: http://ijcmph.com/index.php/ijcmph/article/view/493
11. Date AP, Borkar AS, Badwaik RT, Siddiqui RA, Shende TR, Dashputra A V. Item analysis as tool to validate multiple choice question bank in pharmacology. Int J Basic Clin Pharmacol. 2019;8(9). Available from: https://www.ijbcp.com/index.php/ijbcp/article/view/3324
12. Kaur M, Singla S, Mahajan R. Item analysis of in use multiple choice questions in pharmacology. Int J Appl Basic Med Res. 2016;6(3). Available from: http://www.ijabmr.org/text.asp?2016/6/3/170/186965
13. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Vol. 2, International journal of medical education. 2011. p. 53–5. Available from: http://www.ijme.net/archive/2/cronbachs-alpha/
14. Tavakol M, Dennick R. Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. Med Teach. 2012 Mar;34(3). Available from: http://www.tandfonline.com/doi/full/10.3109/0142159X.2012.651178.
15. Feldt LS. A test of the hypothesis that cronbach's alpha or kuder-richardson coefficent twenty is the same for two tests. Psychometrika. 1969;34(3). Available from: https://doi.org/10.1007/BF02289364
16. Swinscow TDV, Campbell MJ. Statistics at square one. Bmj London; 2002. Available from: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one.
17. Rao C, Prasad H, Sajitha K, … HPIJ of, 2016 undefined. Item analysis of multiple choice questions: Assessing an assessment tool in medical students. ijeprjournal.org [Internet]. [cited 2023 Jul 26]; Available from: https://www.ijeprjournal.org/oaccess.asp?2016/2/4/201/189670/1/1
18. Kheyami D, Jaradat A, Al-Shibani T, Ali FA. Item analysis of multiple choice questions at the department of paediatrics, Arabian gulf university, Manama, Bahrain. Sultan Qaboos Univ Med J. 2018;18(1). Available from: https://journals.squ.edu.om/index.php/squmj/article/view/2525.
19. Bhat SK, Prasad KHL. Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology: A cross-sectional study. Indian J Ophthalmol. 2021;69(2). Available from: https://journals.lww.com/ijo/Fulltext/2021/02000/Item_analysis_and_optimizing_multiple_choice.31.aspx.
20. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. J Pak Med Assoc. 2012 Feb;62(2):142–7. Available from: https://pubmed.ncbi.nlm.nih.gov/22755376/.
21. Bhattacherjee S, Mukherjee A, Bhandari K, Rout A. Evaluation of Multiple-Choice Questions by Item Analysis, from an Online Internal Assessment of 6th Semester Medical Students

in a Rural Medical College, West Bengal. Indian J Community Med [Internet]. 2022 Jan 1 [cited 2023 Jul 14];47(1):92–5. Available from: https://pubmed.ncbi.nlm.nih.gov/35368481/

22. Karelia BN. The levels of difficulty and discrimination indices and relationship between them in four-response type multiple choice questions of pharmacology summative tests of Year II M.B.B.S students. International e-Journal of Science, Medicine & Education. 2013;7(2). Available from: https://doi.org/10.56026/imu.7.2.41
23. Pande SS, Pande SR, Parate VR, Nikam AP, Agrekar SH. Correlation between difficulty & discrimination indices of MCQs in formative exam in Physiology. South-East Asian Journal of Medical Education. 2013;7(1). Available from: https://seajme.sljol.info/article/10.4038/seajme.v7i1.149/
24. Mitra NK. The Levels Of Difficulty And Discrimination Indices In Type A Multiple Choice Questions Of Pre-clinical Semester 1 Multidisciplinary Summative Tests. International e-Journal of Science, Medicine & Education. 2009;3(1). Available from: 10.56026/imu.3.1.2.
25. Gajjar S, Sharma R, Kumar P, Rana M. Item and Test Analysis to Identify Quality Multiple Choice Questions (MCQs) from an Assessment of Medical Students of Ahmedabad, Gujarat. Indian J Community Med [Internet]. 2014 Jan [cited 2023 Jul 14];39(1):17–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24696535