*Medical Education*

# Post-validation item analysis to assess the validity and reliability of multiple-choice questions at a medical college with an innovative curriculum

AMAR IBRAHIM OMER YAHIA

## ABSTRACT

***Background.*** In medical education, the need to obtain reliable and valid assessments is critical for the learning process. This study implemented a post-validation item analysis to create a supply of valid questions for incorporation into the question bank.

***Methods.*** A cross-sectional study was performed in the College of Medicine, University of Bisha, Saudi Arabia. The study was targeting 250 items and 750 distractors from 2017 to 2020. The post-validation item analysis was done to evaluate the quality of the items using test-scoring and reporting software. Data were analysed by SPSS Version 25. Quantitative variables were expressed as mean (SD), while qualitative variables were expressed as number and percentage. An independent *t*-test was done to reveal the association between the item analysis parameters. A value of $p \leq 0.05$ was considered statistically significant.

***Results.*** The mean difficulty index (DIF I), discrimination index (DI) and distractors efficacy (DE) were 73.8, 0.26 and 73.5%, respectively. Of 250 items, 38.8% had an acceptable DIF I (30%–70%) and 66.4% had 'good to excellent' DI (>0.2). Of 750 distractors, 33.6%, 37%, 20% and 9.2% had zero, one, two and three non-functional distractors, respectively. The mean Kuder–Richardson was 0.76. The DIF I was significantly associated with DE (p=0.048). The post-validation item analysis of this study showed that a considerable proportion of questions had acceptable parameters and were recommended for item banking. However, some questions needed to be rephrased and reassessed or discarded.

***Conclusion.*** Three-option multiple-choice questions should be considered for future examinations to improve the assessment process.

Natl Med J India 2021;34:359–62

## INTRODUCTION

An assessment is an essential component of learning.[1] Multiple-choice questions (MCQs) are essential tools used to evaluate the achievement of medical students in various phases during medical education.[2] MCQs are increasingly used in examinations due to their objectivity, ability to test a wide range of content, comparability and limitation of bias by minimizing an individual's judgement during scoring.[3] In medical education, the need to obtain reliable and valid assessments is critical for the learning process. Type-A four-option MCQs are composed of a stem (question), key (best answer) and three other options (the distractors).[4] If the key and distractors given in the question are not standardized, the question will be difficult to answer by candidates or will push the candidates towards a key answer or towards guessing.[5] Item analysis is defined as a process of analysing examinees' responses to evaluate the quality of examination items.[6] The main parameters of item analysis included the difficulty index (DIF I), discrimination index (DI), distractors efficacy (DE) and Kuder–Richardson (KR-20) formula. DIF I reflects the ratio of examinees who correctly respond to the questions. DI measures the item's ability to differentiate between high and low achievers. DE measures how many alternatives, other than the key, are distracting the students from choosing the key answer.[7] KR-20 is the item analysis parameter that determines the reliability of the examination. According to the medical education policy, the ideal MCQs should have a DIF I of 30%–70%, a DI of $\geq 0.2$, a DE of 100% and KR-20 from 0 to 1.[8,9]

Item analysis determines whether the question should be stored, rephrased or discarded.[10] It also provides feedback for the examiners to modify their questions so they are more valid and reliable for the next assessment.[11,12] Well-constructed MCQs assess higher cognitive domains of Bloom's taxonomy and differentiate the examinees' different skills.[13] Effective distractors are essential for constructing ideal MCQs. The distractors must be constructed based on a common misconception about the answer.[14] Functional distractors (FDs) are options chosen by at least 5% of candidates, while non-FDs (NFDs) are options chosen by <5% of the examinees.[8,9] Studying the functional status of the items is of interest as a plausible framing distractor improves the test quality.[15] Improving item quality is possible by removing the item flaws and analysing the items' real performance. This study implemented a post-validation item analysis to create a supply of valid questions for incorporation into the question bank.

## METHODS

### *Study design and setting*

A cross-sectional study was conducted at the College of Medicine, University of Bisha, Saudi Arabia, from November 2019 to March 2020. The College of Medicine, University of Bisha is a 6-year-old college adopting a student-centred approach during the process of learning (innovative curriculum).

College of Medicine, University of Bisha, PO Box 199, Bisha 61922, Saudi Arabia
AMAR IBRAHIM OMER YAHIA Department of Basic Medical Sciences, Unit of Pathology (Haematology)
Correspondence to AMAR IBRAHIM OMER YAHIA; *amarfigo2@yahoo.com*

**[To cite:** Yahia AI. Post-validation item analysis to assess the validity and reliability of multiple-choice questions at a medical college with an innovative curriculum. *Natl Med J India* 2021;**34:**359–62.]

### *Data collection and procedure*

The study targeted 250 items and 750 distractors from the examination of the haematology course administered to level five medical students for the past three consecutive academic years (2017–2020). The MCQs were constructed by subject experts according to the guideline references using an examination blueprint that aligns each item with the corresponding specific learning outcome. As per the college regulation, the examination should be approved by the student assessment committee (SAC) before utilization. The SAC is responsible for student assessments, approving of examination blueprints, examination questions and results. The SAC members are faculty experts in assessment. The SAC policies are regularly reviewed and updated. According to the SAC policy, all MCQs are newly constructed, the correct answer was allotted one mark with no negative score for the wrong response and a passing grade is 60%. The examination comprised type-A MCQs with a single best answer. Each MCQ had a stem and four options, one key (correct answer) and three distractors (incorrect responses). Possible copying from neighbouring students was avoided by appointing two invigilators with cameras in the assessment hall and ensuring a reasonable distance between students. In addition, we administered one different examination paper model for every 10 students according to the SAC policy.

### *Post-validation item analysis*

The post-validation item analysis was performed to evaluate the quality of the items using the Apperson DataLink 3000 Scanner Kit (test-scoring and reporting software). According to the results of item analysis, the SAC stored, rephrased or removed the question. The DIF I ranged from 0% to 100%. The criteria of categorization for DIF I were: a DIF >70% indicated that the item was easy, a DIF of 30%–70% was acceptable and a DIF <30% indicated a difficult item. The range of DI is 0–1. A DI <0.2 indicates a poor item, a DI of 0.2–0.34 indicates a good DI and a DI ≥0.35 indicates an excellent item. The DE range in four-option MCQs was from 0% to 100%. Items with three, two, one and zero NFDs had a DE of 0%, 33%, 66% and 100%, respectively.[16] A KR-20 below 0.7 indicated poor examination reliability, while a KR-20 equal to or above 0.7 was considered acceptable. Questions with DIF I and DI out of the acceptable ranges usually had a low KR-20 value.[17]

### *Data analysis*

Data were analysed by IBM SPSS Statistics for Windows, Version 25. Quantitative variables were expressed as mean (SD), while qualitative variables were expressed as a number and percentage. An independent *t*-test was done to reveal the association between the item analysis parameters. A p value <0.05 was considered statistically significant. Items were classified as excellent, acceptable or poor based on the results of DIF I, DI and DE and decisions to store, rephrase or discard were recommended accordingly. This study did not deal with human subjects, so an ethical review by the Institutional Review Board was not sought. However, the examination office at the institute allowed access to the study data.

## RESULTS

A total of 250 MCQs and 750 distractors were analysed. The number of examinees ranged from 43 to 48, while the number of questions ranged from 60 to 100. The mean scores of students ranged from 65.7% to 75.4%, and the KR-20 ranged from 0.71 to 0.85 (Table I). The mean (SD) DIF I was 73.8 (23.7), the mean (SD) DI was 0.26 (0.2) and the mean (SD) DE was 73.5 (23.7). Of the 250 items, 97 (38.8%) had a DIF I of 30%–70%, while 153 (61.2%) had a DIF I of <30% or >70%. Approximately two-thirds (66.4%) of the total items had a DI of ≥0.2, while 63 (25.2%) of the items had a DI of <0.2%. In addition, 21 (8.4%) items had a negative DI (Table II). Of 750 distractors, 488 (65.1%) were functioning effectively, with their DE being 100%. The proportion of items containing zero, one, two and three NFDs was 33.6%, 37%, 20%

TABLE I. Characteristics of the examinations, 2017–2020

| Characteristic | 2017–2018 | 2018–2019 | 2019–2020 |
|---|---|---|---|
| Number of items | 60 | 100 | 90 |
| Number of examinees | 43 | 46 | 48 |
| Mean test score (%) | 67.0 | 65.7 | 75.4 |
| Range of test score (%) | 36.3–92.5 | 32.4–91.7 | 43.3–93.3 |
| Kuder–Richardson-20 | 0.71 | 0.85 | 0.72 |

TABLE II. Distribution of items in relation to difficulty index I, discrimination index and actions proposed

| Categorization | Number of items (%) | Interpretation | Action |
|---|---|---|---|
| *Difficulty index* | | | |
| 30–70 | 97 (38.8) | Good/acceptable | Store |
| <30 | 16 (6.4) | Difficult | Revise/discard |
| >70 | 137 (54.8) | Easy | Modification/ discard |
| *Discrimination index* | | | |
| ≥0.35 | 98 (39.2) | Excellent | Store |
| 0.2–0.34 | 68 (27.2) | Good/acceptable | Store |
| 0–0.19 | 63 (25.2) | Poor | Modification/ discard |
| <0 | 21 (8.4) | Bad | Discard |

TABLE III. Frequency distribution of non-functional distractors according to selection

| Distractor analysis | *n* (%) |
|---|---|
| Items with 0 NFD (DE=100) | 84 (33.6) |
| Items with 1 NFD (DE=66) | 93 (37) |
| Items with 2 NFDs (DE=33) | 50 (20) |
| Items with 3 NFDs (DE=0) | 23 (9.2) |
| Overall mean (SD) DE | (73.5 [23.7]) |

NFD non-functional distractor DE distractors efficacy

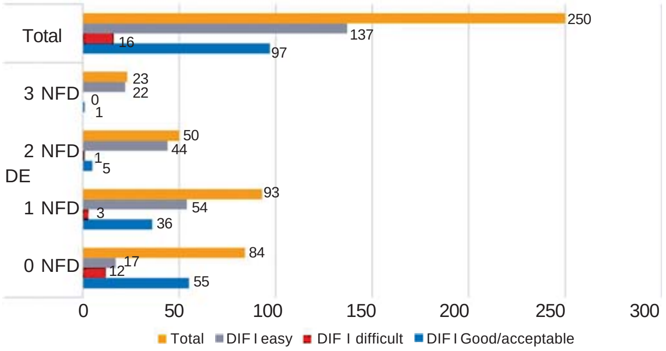

FIG 1. Correlation between difficulty index (DI) and distractors efficacy (DE)

and 9.2%, respectively (Table III). The DIF I was significantly associated with the DE (p=0.048; Fig. 1). Other item analysis parameters showed no significant association (p>0.05).

## DISCUSSION

Post-examination evaluation of MCQs through item analysis is an essential process for identifying ideal questions and developing a valid pool of MCQs for future assessment. Item analysis assists in recognizing the quality of questions that determines the quality of the assessment and allows for further improvement and development.[18] In the present study, 250 items of type-A MCQs with 750 distractors from the haematology course examination were evaluated to determine the DIF I, DI, DE and KR-20. Ideal MCQs should have DIF I of 30%–70% with DI of at least 0.2 and 100% DE.[8] In our study, the mean (SD) DIF I was 73.8 (23.7)%, with 38.8% of the items within the acceptable range (30%–70%), 6.4% of the items were difficult (<30%) and 54.8% of the items were easy (>70%). A study conducted in a medical institute in India evaluated 200 MCQs and found that 46% of the questions were in the average range, 37% were difficult and 16% were easy.[19] Another study by Patil *et al.* showed that 46% of items fall under the classification of good, 17% were easy and 37% were difficult.[20] Karelia *et al.* reported a range of mean (SD) between 47.17 (19.77) and 58.08 (19.33) and reported 61% of the items fell in the acceptable range, 24% of the items were easy and 15% of the items were difficult.[21] Other studies have proposed the mean of DIF I as 48.9, 52.5, 57.7 and 62%.[22–25] However, Mitra reported that the mean DIF I was between 64% and 89% among 12 examinations taken between 2003 and 2006.[26] The SAC policy in our institute determined that the DIF I was between 25% and 85%, which might explain why the mean DIF I in our study was high (easy items) in comparison to other studies. In addition, it might be due to the high proportion of items that did not have any functioning distractors (34.9%). Increasing the number of easy items will lead to inflated marks and result in declining motivation, while an increasing number of difficult items will result in deflated scores.[7] Furthermore, too easy and difficult items will result in poor DI. Despite this, easy MCQs can be kept and placed at the beginning of the assessment to raise the confidence of examinees while the hard items should be placed at the end of the assessment to differentiate between good and poor candidates.

Difficult questions may indicate that the test item is not delivered correctly or the scientific content is tough to understand.[26] Difficult questions must be checked for possible confusing language, contents of controversies, under-coverage of scientific materials, inappropriate difficulty level or an incorrect key. The current study showed that the mean (SD) was 0.26 (0.2), which indicates good discriminating items (DI $\geq$0.2). These findings are consistent with the study carried out by Patel[27] and are higher in comparison to the study of Gajjar *et al.* (DI=0.14 [0.19]).[13] In our study, 98 (39.2%) of the questions showed an excellent predisposition to distinguish candidates of upper and lower marks (D $\geq$0.35), while 68 (27.2%) and 63 (25.2%) MCQs demonstrated good (DI=0.2–0.34) and poor (DI<0.2) discrimination ability, respectively. This finding of our present study is comparable to the literature.[19] In another study, 46% of the questions had an excellent DI, 32% of the items had a good/acceptable DI and 22% of the items had a poor DI.[21] Mehta and Mokhasi found that the items with a DI >0.35 were 10 (50%), items with a DI between 0.2 and 0.34 were 4 (20%) and items with a DI <0.2 were 6 (30%).[22] In the current study, 21 (8.4%) questions showed negative DI values. Some studies have shown a negative DI in 20% of the items.[13] A negative DI usually can be explained by the wrong key, vague wordings or unclear areas of under-standing.[28]

In addition to the mentioned reasons, unclear item construction and non-readiness for the assessment are also explanations for a negative DI. Items that failed to differentiate adequately between the candidates should be evaluated for possible flaws and rephrased or discarded accordingly. The negative DI reflects negatively on the assessment validity.[10] I recommended the follow-up by investigating the key and any possible technical flaw by the SAC; the item can then be stored after the correction of the key or rephrasing the question if any. The current study indicates that the mean (SD) DE was 73.5 (23.7)%. This finding is almost similar to the finding reported by Gajjar *et al.*[13] but lower than in the study reported by Hingorjo and Jaleel.[8] Furthermore, it is not in line with the present-day study conducted by Patel, in which the DE was 84.9.[27]

In our study, among 750 distractors, 488 (65.1%) were FDs, while 262 (34.9%) were NFDs; this finding is comparable to the findings of Haladyna and Downing (38% NFDs).[29] Gajjar *et al.* concluded that in a total of 150 distractors, 133 (89.6%) were FDs, while 17 (11.4%) were NFDs.[13] This result indicates the difficulty of constructing plausible distractors in four- and five-option MCQs, particularly in assessing the knowledge of content. In addition, item writers sometimes find difficulty in developing plausible distractors, and some distractors are just for completing the options. Three-option MCQs should be considered for future examinations when there is difficulty in modifying the four- or five-option MCQs, and there is no other assessment method apart from the MCQs. Regarding the details of the functional status of the item distractors, in one-third of the items (33.6%), all of the distractors were sufficiently attractive to be selected (zero NFD with 100% DE), while 93 (37.2%) and 50 (20%) had one and two FDs, respectively. Only 23 (9.2%) MCQs had no FDs. Our study found a higher proportion of NFDs as compared to other studies conducted by Kolte[30] and Patil *et al.*[20] On the contrary, the proportion of items containing all three functioning distractors in the current study was higher than the percentage reported by Tarrant *et al.*[9] and Sayyah *et al.*[31] Another study conducted by Sharif *et al.* showed that 15.3%, 38.1% and 34.6% of the questions had three, two and one NFD, respectively, whereas 12% of the items had zero NFDs as similar to the present study.[2] Haladyna and Downing reported that approximately two-thirds of the MCQs they evaluated had two or only one FD and none had a DE of 100%.[29] This finding might be due to MCQs constructed by doctors from the hospital participating as a part-time instructor with less experience in the construction of an ideal MCQ. NFDs should either be removed or replaced with more plausible options. Questions assessing factual knowledge can be kept, but the passing score should be set according to the standard setting. There is no scientific justification that all MCQs should have an equal number of distractors. The number of distractors should depend on the options according to the content area being assessed, but unfortunately, institutional guidelines sometimes restrict this. Designing plausible distractors to reduce NFDs is a critical part of constructing good MCQs. More or less NFDs in the question will affect the DIF I and have discriminative power. Items with three or two FDs were significantly harder to answer than questions with one or no FDs. The mean KR-20 in the present study was 0.76, which indicates a reliable examination and assessment.

Examinations with a higher number of items are more reliable. Our result found a significant association between the DE and DIF I, signifying that items with a higher DE were more difficult to answer. Overall, in this study, as per the DIF I, DI and DE, there were a total of 55 (22%) items that fulfilled the characterization of ideal items. Those ideal questions are appropriate for item banking for future utilization. Of concern are some items that needed to be rephrased and reassessed or discarded. Item reassessing is a continuous process and should be done frequently to optimize the quality of assessment items. Questionable items were discussed in the SAC meeting with concerned faculty members, and any required modifications were done to improve the questions.

### *Strength and limitations*

To our knowledge, this is the first study that analysed the items of the haematology course in our institute. The present study included 250 items and 750 distractors, which was considered a good number for item evaluation.

The results observed in this study addressed only one course and did not reflect other courses. This study was performed in a country where English is not the first language for the examinees so that may affect their ability to respond to questions.

### *Conclusion*

Constructing ideal questions is an effective way to improve the validity and reliability of an examination. Item analysis is an essential tool serving as an effective feedback mechanism for improvement of questions. The post-validation item analysis of this study showed that a considerable proportion of questions had acceptable parameters and were recommended for item banking; however, some questions needed to be rephrased and reassessed or discarded. Discussing the results of item analysis with the faculty members and students helped to improve the educational assessment. The results of the current study concluded that item writers had difficulty in developing plausible distractors. Hence, three-option MCQs should be considered for future examinations to improve the assessment process. Regular faculty development programmes and workshops in item construction should be offered to the faculty members to improve their skills in constructing ideal MCQs. Cyclic review of the questions in the question bank after each examination to identify the areas needing revision and update is recommended.

## ACKNOWLEDGEMENT

The author would like to thank the examination office at the College of Medicine, University of Bisha, for providing the study material.

*Conflicts of interest.* None declared

## REFERENCES

1 Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: An unavoidable truth? *Anat Sci Educ* 2009;**2:**199–204.

2 Sharif M, Rahimi SM, Rajabi M, Sayyah M. Computer software application in item analysis of exams in a college of medicine. *ARPN J Sci Tech* 2014;**4:**565–9.

3 Loh KY, Elsayed I, Nurjahan MI, Roland GS. Item difficulty and discrimination index in single best answer MCQ: End of semester examinations at Taylor's clinical school. In: *Redesigning learning for greater social impact*. Singapore:Springer; 2018:167–71.

4 Cizek GJ, O'Day DM. Further investigation of non functioning options in multiple-choice test items. *Educ Psychol Measurement* 1994;**54:**861–72.

5 Kuechler WL, Simkin MG. Why is performance on multiple choice tests and constructed response tests not more closely related? Theory and an empirical test. *Dec Sci J Innovat Educ* 2010;**8:**55–73.

6 Singh T, Gupta P, Singh D. Test and item analysis. In: *Principles of medical education*. 4th ed. New Delhi:Jaypee Brothers Medical Publishers; 2013:109.

7 Eaves S, Erford B. The Gale group: The purpose of item analysis, item difficulty, discrimination index, characteristic curve. Available at *www.education.com/reference/article/itemanalysis/* (accessed on 15 Apr 2013).

8 Hingorjo MR, Jaleel F**.** Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *J Pak Med Assoc* 2012;**62:**142–7.

9 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:**40.

10 Matlock-Hetzel S. Basic concept in item and test analysis. Available at *http://ericae.net/ft/tamu/Espy.html* (accessed on 15 Apr 2018).

11 Namdeo SK, Sahoo B. Item analysis of multiple choice questions from an assessment of medical students in Bhubaneswar, India. *Int J Res Med Sci* 2016;**4:**1716–19.

12 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 Singapore* 2006;**35:**67–71.

13 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* 2014;**39:**17–20.

14 Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Measurement Educ* 2002;**15:**309–33.

15 Bruno JE, Dirkzwager A. Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educ Psychol Measurement* 1995;**55:**959–66.

16 Chauhan PR, Ratrhod SP, Chauhan BR, Chauhan GR, Adhvaryu A, Chauhan AP. Study of difficulty level and discriminating index of stem type multiple choice questions of anatomy in Rajkot. *Biomirror* 2013;**4:**1–4.

17 Panchal P, Prasad B, Kumari S. Multiple choice questions—role in assessment of competency of knowledge in anatomy. *Int J Anat Res* 2018;**6:**5156–62.

18 Liu NF, Carless D. Peer feedback: The learning element of peer assessment. *Teach Higher educ* 2006;**11:**279–90.

19 Christian DS, Prajapati AC, Rana BM, Dave VR. Evaluation of multiple choice questions using item analysis tool: A study from a medical institute of Ahmedabad, Gujarat. *Int J Community Med Public Health* 2017;**4:**1876.

20 Patil R, Palve SB, Vell K, Boratne AV. Evaluation of multiple choice questions by item analysis in a medical college at Pondicherry, India. *Int J Community Med Public Health* 2016;**3:**1612–16.

21 Karelia BN, Pillai A, Vegada 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 MBBS students. *IeJSME* 2013;**7:**41–6.

22 Mehta G, Mokhasi V. Item analysis of multiple choice questions—An assessment of the assessment tool. *Int J Health Sci Res* 2014;**4:**197–202.

23 Kaur M, Singla S, Mahajan R. Item analysis of in use multiple choice questions in pharmacology. *Int J Appl Basic Med Res* 2016;**6:**170.

24 Pande SS, Pande SR, Parate VR, Nikam AP, Agrekar SH. Correlation between difficulty and discrimination indices of MCQs in formative exam in Physiology. *South East Asian J Med Educ* 2013;**7:**45–50.

25 Chauhan P, Chauhan GR, Chauhan BR, Vaza JV, Rathod SP. Relationship between difficulty index and distracter effectiveness in single best-answer stem type multiple choice questions. *Int J Anatomy Res* 2015;**3:**1607–10.

26 Mitra NK, Nagaraja HS, Ponnudurai G, Judson JP. The levels of difficulty and discrimination indices in type a multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. *IeJSME* 2009;**3:**2–7.

27 Patel RM. Use of item analysis to improve quality of multiple choice questions in II MBBS. *J Educ Technol Health Sci* 2017;**4:**22–9.

28 Bauer D, Kopp V, Fischer MR. Answer changing in multiple choice assessment change that answer when in doubt-and spread the word! *BMC Med Educ* 2007;**7:**28.

29 Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? *Educ Psychol Measurement* 1993;**53:**999–1010**.**

30 Kolte V. Item analysis of multiple choice questions in physiology examination. *Indian J Basic Appl Med Res* 2015;**4:**320–6.

31 Sayyah M, Vakili Z, Masoudi Alavi N, Bigdeli M, Soleymani A, Assarian M, *et al*. An item analysis of written multiple-choice questions: Kashan University of Medical Sciences. *Nurs Midwifery Stud* 2012;**1:**83–7.