Sultan Qaboos University Med J, February 2018, Vol. 18, Iss. 1, pp. e68–74, Epub. 4 Apr 18
Submitted 30 Aug 17
Revisions Req. 22 Oct & 12 Dec 17; Revisions Recd. 12 Nov & 25 Dec 17
Accepted 7 Jan 18
doi: 10.18295/squmj.2018.18.01.011

CLINICAL & BASIC RESEARCH

# Item Analysis of Multiple Choice Questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain

Deena Kheyami,[1] Ahmed Jaradat,[2] Tareq Al-Shibani,[3] *Fuad A. Ali[1]

# تحليل مفردات أسئلة الاختيار من متعدد في قسم الأطفال بجامعة الخليج العربي، المنامة، البحرين

دينا خيامي، أحمد جردات، طارق الشيباني، فؤاد عبدالله علي

**abstract:** ***Objectives:*** The current study aimed to carry out a post-validation item analysis of multiple choice questions (MCQs) in medical examinations in order to evaluate correlations between item difficulty, item discrimination and distraction effectiveness so as to determine whether questions should be included, modified or discarded. In addition, the optimal number of options per MCQ was analysed. ***Methods:*** This cross-sectional study was performed in the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain. A total of 800 MCQs and 4,000 distractors were analysed between November 2013 and June 2016. ***Results:*** The mean difficulty index ranged from 36.70–73.14%. The mean discrimination index ranged from 0.20–0.34. The mean distractor efficiency ranged from 66.50–90.00%. Of the items, 48.4%, 35.3%, 11.4%, 3.9% and 1.1% had zero, one, two, three and four nonfunctional distractors (NFDs), respectively. Using three or four rather than five options in each MCQ resulted in 95% or 83.6% of items having zero NFDs, respectively. The distractor efficiency was 91.87%, 85.83% and 64.13% for difficult, acceptable and easy items, respectively ($P$ <0.005). Distractor efficiency was 83.33%, 83.24% and 77.56% for items with excellent, acceptable and poor discrimination, respectively ($P$ <0.005). The average Kuder-Richardson formula 20 reliability coefficient was 0.76. ***Conclusion:*** A considerable number of the MCQ items were within acceptable ranges. However, some items needed to be discarded or revised. Using three or four rather than five options in MCQs is recommended to reduce the number of NFDs and improve the overall quality of the examination.

***Keywords:*** Medical Education; Educational Measurement; Academic Performance; Psychometrics; Examination Questions; Discriminant Analysis; Bahrain.

**الملخص: الهدف:** هدفت الدراسة الحالية إلى إجراء تحليل مفردات أسئلة الاختيار من متعدد ودراسة العلاقة بين مؤشر الصعوبة، مؤشرالتمييز وفعالية المشتتات من أجل الإحتفاظ بالأسئلة أو تعديل أو إلغاء كل سؤال. بالإضافة إلى ذلك تم تحليل أفضل عدد من البدائل في كل سؤال. **الطريقة:** هذه دراسة مستعرضة أجريت في قسم طب الأطفال بجامعة الخليج العربي، المنامة، البحرين. تم تحليل 800 سؤال إختيار من متعدد و عدد 4,000 من المشتتات في الفترة من نوفمبر 2013 الى يونيو 2016. **النتائج:** تراوح متوسط مؤشر الصعوبة بين 36.70–73.14% و تراوح مؤشر التمييز من 0.20–0.34 و تراوحت فعالية المشتتات من 66.50–90.00% على التوالي. بلغت نسبة الفقرات التي تحتوي على صفر، واحد، اثنان، ثلاثة وأربعة مشتتات غير فاعلة: 48.4%، 35.3%، 11.4%، 3.9% و 1.1% على التوالي. باستخدام ثلاثة أو أربعة بدلا من خمسة بدائل وبالتالي إزالة واحد أو اثنين من المشتتات الغير فاعلة من شأنه أن تكون 95% أو 83.6% من الاسئلة تحوي على صفر من المشتتات الغير فاعلة على التوالي. بلغت فعالية المشتتات 91.87%، 85.83% و 64.13% بالنسبة للأسئلة ذات مؤشر الصعوبة في المعدلات الصعبة والجيدة والسهلة على التوالي ($P$ <0.005). وبلغت فعالية المشتتات 83.33%، 83.24% و 77.56% بالنسبة للأسئلة ذات مؤشر التمييز في المعدلات الممتازة والجيدة والضعيفة على التوالي ($P$ <0.005). كان متوسط اختبار الموثوقية 20 ب كودر ريتشارسون هو 0.76. **الخلاصة:** كانت نسبة جيدة من مفردات أسئلة الاختيار من متعدد ضمن المعدلات المقبولة. ومع ذلك، يلزم إستبعاد بعض الفقرات أو تنقيحها. نوصي بإستخدام ثلاثة أو أربعة بدلا من خمسة بدائل للحد من المشتتات الغير فاعلة ولتحسين جودة أسئلة الاختيار من متعدد.

**الكلمات المفتاحية:** التعليم الطبي؛ القياس التربوي؛ الأداء الأكاديمي؛ السيكومترية؛ أسئلة الأمتحانات؛ تحليل التمييز؛ البحرين.

## Advances in Knowledge

- *Designing adequate multiple choice questions (MCQs) is essential to assess learning among medical students. Item analysis is an important scientific tool that provides information about the reliability and validity of MCQ items. However, item analysis studies are limited, particularly in medical schools in Arabian Gulf countries.*
- *The findings of the current study will hopefully increase awareness of this measurement tool among medical education providers in the region.*

## Application to Patient Care

- *Designing appropriate MCQs improves the assessment and learning output of medical students. High-quality medical education in the Arabian Gulf region will encourage the provision of enhanced healthcare services to local populations.*

*Departments of [1]Paediatrics, [2]Family & Community Medicine and [3]Physiology, Arabian Gulf University, Manama, Bahrain*
**Corresponding Author's e-mail: fuadama@agu.edu.bh*

While assessment is an essential part of student learning, assessment tools need to be valid, reliable and objective and reflect various achievement levels. Multiple choice questions (MCQs) should not only aim to assess knowledge recollection, but also measure other teaching objectives within Bloom's taxonomy of learning, such as comprehension, application, analysis, synthesis and evaluation.[1] Constructing a high-quality MCQ examination can be difficult and time-consuming; however, this approach is usually preferential to other types of assessment tools because it is objective and leaves little room for human bias, as answers to MCQ questions can be easily and reliably scored.[2,3] In recent years, the most common type of MCQs employed in examinations are type A MCQs, which consist of a stem followed by four or five options or distractors.[4,5]

An item analysis assesses the reliability and validity of an examination by examining student performance with regards to each MCQ and applying statistical analyses to determine whether the item should be kept, reviewed or discarded from the test. Common item analysis parameters include the difficulty index (DIFI), which reflects the percentage of correct answers to total responses; the discrimination index (DI), also known as the point biserial correlation, which identifies discrimination between students with different levels of achievement; and distractor efficiency (DE), which indicates whether the distractors in the item are well-chosen or have failed to distract students from selecting the correct answer. An ideal item should have a DIFI of between 30–70%, a DI of >0.2 and a DE of 100%.[6,7]

At the end of their 10-week clinical rotation in the Department of Paediatrics of the Arabian Gulf University (AGU), Manama, Bahrain, paediatric clerkship students undergo MCQ examinations in addition to objective standard clinical examinations, short-answer question tests and continuous performance assessments. Each MCQ consists of a stem followed by five distractors. Students do not receive negative marks for wrong answers and the tests are criterion-referenced, with passing standards expressed in absolute terms and a passing score of 60%. For each examination, approximately 50% of the MCQs are newly constructed while the remaining questions are taken from a question bank after revising and modifying question items according to item analysis outcomes. However, the examinations are not assessed for equivalent difficulty across the years.

The current study aimed to carry out a post-validation item analysis of MCQs used in end-of-rotation examinations between 2013–2016 at the AGU Department of Paediatrics. Based on the item analysis outcomes, recommendations were made as to whether the questions should be retained, modified or discarded from the AGU question bank. In addition, correlations between the difficulty, item discrimination and distraction effectiveness of each item were calculated and the optimal number of options in each MCQ was determined.

## Methods

This cross-sectional study was performed in the Department of Paediatrics at AGU and included all MCQ items of paediatric clerkship summative examination papers from November 2013 to June 2016. There were 50 MCQs per paper and four examinations per year, resulting in a total of 800 MCQs and 4,000 distractors. In total, 608 students had taken the exam-

**Table 1:** Mean difficulty index, discrimination index and distractor efficiency of end-of-rotation paediatric examinations at the Arabian Gulf University, Manama, Bahrain (N = 16)

| Year | Exam. | Mean ± SD | | |
|---|---|---|---|---|
| | | DIFI % | DI | DE % |
| 2013 | 1 | 65.81 ± 24.00 | 0.34 ± 0.22 | 70.00 ± 28.12 |
| | 2 | 73.14 ± 20.39 | 0.30 ± 0.16 | 68.00 ± 29.47 |
| | 3 | 70.70 ± 19.34 | 0.30 ± 0.20 | 66.50 ± 28.84 |
| | 4 | 58.72 ± 23.19 | 0.23 ± 0.19 | 79.00 ± 21.64 |
| | Total | 67.09 ± 22.34 | 0.29 ± 0.20 | 70.88 ± 27.43 |
| 2014 | 5 | 56.06 ± 23.38 | 0.28 ± 0.17 | 83.50 ± 19.31 |
| | 6 | 52.40 ± 18.74 | 0.28 ± 0.27 | 75.00 ± 23.15 |
| | 7 | 51.70 ± 24.21 | 0.20 ± 0.27 | 83.00 ± 17.81 |
| | 8 | 52.88 ± 21.20 | 0.29 ± 0.20 | 85.00 ± 16.75 |
| | Total | 53.26 ± 21.88 | 0.26 ± 0.23 | 81.62 ± 19.65 |
| 2015 | 9 | 39.54 ± 21.25 | 0.28 ± 0.19 | 90.00 ± 15.15 |
| | 10 | 44.78 ± 23.63 | 0.23 ± 0.18 | 88.50 ± 20.34 |
| | 11 | 43.59 ± 22.46 | 0.27 ± 0.19 | 86.50 ± 16.91 |
| | 12 | 44.73 ± 25.25 | 0.23 ± 0.18 | 81.00 ± 21.17 |
| | Total | 43.16 ± 23.12 | 0.25 ± 0.18 | 86.50 ± 18.73 |
| 2016 | 13 | 41.84 ± 22.07 | 0.27 ± 0.15 | 85.50 ± 18.96 |
| | 14 | 36.70 ± 23.26 | 0.22 ± 0.18 | 89.00 ± 16.87 |
| | 15 | 54.68 ± 18.94 | 0.31 ± 0.14 | 84.00 ± 20.05 |
| | 16 | 47.14 ± 22.84 | 0.24 ± 0.15 | 89.00 ± 15.29 |
| | Total | 45.09 ± 22.68 | 0.26 ± 0.18 | 86.88 ± 17.89 |
| **Average** | | **52.15 ± 24.37** | **0.27 ± 0.20** | **81.47 ± 22.19** |

*Exam. = examination; SD = standard deviation; DIFI = difficulty index; DI = discrimination index; DE = distractor efficiency.*

**Table 2:** Non-functioning distractors per individual multiple choice question items in the end-of-rotation paediatric examinations at the Arabian Gulf University, Manama, Bahrain (N = 800)

| Year | Parameter | Number of NFDs per item | | | | | Total |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | |
| 2013 | n (%) | 64 (32) | 73 (36.5) | 37 (18.5) | 18 (9) | 8 (4) | 200 (100) |
| | Mean DIFI % | 51.62 | 66.60 | 75.62 | 91.96 | 100.00 | 67.09 |
| | Mean DI | 0.33 | 0.32 | 0.29 | 0.18 | 0.00 | 0.29 |
| 2014 | n (%) | 88 (44) | 84 (42) | 21 (10.5) | 7 (3.5) | 0 (0) | 200 (100) |
| | Mean DIFI % | 45.54 | 56.27 | 65.01 | 78.81 | - | 53.26 |
| | Mean DI | 0.26 | 0.28 | 0.28 | 0.07 | - | 0.26 |
| 2015 | n (%) | 116 (58) | 66 (33) | 13 (6.5) | 4 (2) | 1 (0.5) | 200 (100) |
| | Mean DIFI % | 39.19 | 43.29 | 60.24 | 86.38 | 100.00 | 43.16 |
| | Mean DI | 0.24 | 0.27 | 0.27 | 0.19 | 0.00 | 0.25 |
| 2016 | n (%) | 119 (59.5) | 59 (29.5) | 20 (10) | 2 (1) | 0 (0) | 200 (100) |
| | Mean DIFI % | 38.98 | 48.89 | 65.02 | 97.37 | - | 45.09 |
| | Mean DI | 0.27 | 0.25 | 0.27 | 0.16 | - | 0.26 |
| **Total** | **n (%)** | **387 (48.4)** | **282 (35.3)** | **91 (11.4)** | **31 (3.9)** | **9 (1.1)** | **800 (100)** |
| | **Mean DIFI %** | **42.62** | **54.36** | **68.65** | **88.62** | **100.00** | **52.15** |
| | **Mean DI** | **0.27** | **0.28** | **0.28** | **0.15** | **0.00** | **0.27** |

*NFDs = non-functioning distractors; DI = discrimination index.*

inations during the study period, with an average of 38 students sitting each examination.

Items were only used for summative assessment and were not reviewed with the students at any time. The content and construct validity of the examinations were verified by the Paediatric Examination Committee, which consisted of five content experts and paediatric consultants. Each examination was designed according to a predetermined examination blueprint, ensuring that all essential knowledge and skills were covered based on learning objectives. The post-validation item analysis was performed using the Oracle Database, Version 10g (Oracle Corp., Redwood City, California, USA). The committee discarded existing MCQs based on the item analysis results, flaws in MCQ construction and how frequently an item was used in previous years. The question bank was secured in the assessment office to which only authorised individuals were allowed access via a digital security system. Examinees entered their answers in pencil on a Scantron® optical answer sheet (Scantron Corp., Tustin, California, USA).

The item analysis parameters used in the current study included the DIFI, DI and DE. The DIFI ranged from 0% (i.e. none of the students answered the item correctly) to 100% (i.e. all of the students answered the item correctly). In general, items with a DIFI of <30% were considered difficult, those between 30–70% were considered acceptable and those >70% were considered easy. Kelley's method was used to calculate the DI based on the difference between the scores of high-achievers, classified as the top 27% of test-takers, and low-achievers, classified as the bottom 27% of test takers.[8] The larger the difference between the high- and low-achieving groups, the higher the DI of an item. The DIs of items ranged from -1 (all and only low achievers answered correctly) to +1 (all and only high achievers answered correctly). Items with a DI of ≥0.35 were considered excellent, those between 0.2–0.34 were considered acceptable and those <0.2 were considered poor.

The DE was calculated based on the number of nonfunctional distractors (NFDs) per item. An NFD was defined as an incorrect MCQ option selected by less than 5% of students.[7] The DE was deemed to be either 0%, 25%, 50%, 75% or 100% if an item had four, three, two, one or zero NFDs, respectively.[9] The reliability of the examination was measured using the Kuder-Richardson formula 20 coefficient ($KR_{20}$); this value usually ranges from 0–1, with higher $KR_{20}$ values (i.e. closer to 1) indicating greater reliability. A $KR_{20}$ value of <0.3 is considered poor and a value of ≥0.7 is considered acceptable. Items with DIFIs of >70% or <30% usually yield a low $KR_{20}$ value, as do items with a DI of <0.2.[2,10,11]

Data analysis was performed using the Statistical Package for the Social Sciences (SPSS), Version 23.0

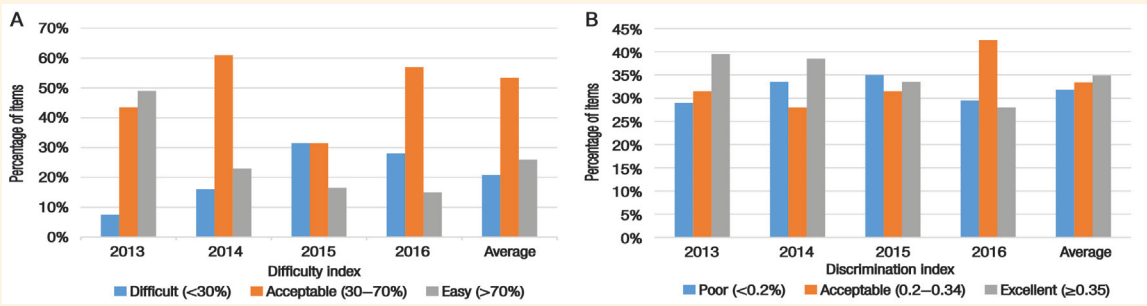


**Figure 1:** Distribution according to (**A**) difficulty index and (**B**) discrimination index of multiple choice questions in end-of-rotation paediatric examinations at the Arabian Gulf University, Manama, Bahrain (N = 800).

(IBM Corp., Armonk, New York, USA). Variables were presented as means ± standard deviations. The linear relationship between DIFI and DI was measured using Pearson's correlation test. A two-way analysis of variance was used to examine the differences in DE (dependent variable), DIFI (independent variable one) and DI (independent variable two). A *P* value of <0.050 was considered statistically significant.

The Vice Dean for Academic Affairs at AGU approved this study and allowed access to the examination data. The identities of the students taking the examination were kept anonymous and confidential. No human participants were involved in this study.

## Results

The mean DIFI of the examinations ranged from 36.70% in 2016 to 73.14% in 2013, with the overall mean DIFI considered acceptable (52.15%). The overall mean DI and DE ranged between 0.20–0.34 and 66.50–90.00%, respectively [Table 1]. Of the total number of items, 48.4%, 35.3%, 11.4%, 3.9% and 1.1% had zero, one, two, three and four NFDs, respectively [Table 2]. It was calculated that using four rather than five options in each MCQ by removing one NFD would result in 83.6% of items having zero NFDs. Using three options and removing two NFDs resulted in 95% of items having zero NFDs.

The overall DIFI increased as the number of NFDs increased, with DIFIs of 42.62%, 54.36%, 68.65%, 88.62% and 100% for items with zero, one, two, three and four NFDs, respectively. This finding was observed in the mean DIFI for each year as well. The overall mean DIFI was almost the same for items with zero, one and two NFDs (0.27%, 0.28% and 0.28%), while they were 0.15% and 0% for items with three and four NFDs, respectively. Similar results were observed for the mean DI of each year as well. More than 40% of the MCQs had an acceptable DIFI throughout the study period. The lowest percentage of difficult MCQs was observed in 2013 (7.5%). The highest percentage of difficult MCQs was 31.5%, noted in 2015. There were more easy MCQs in 2013 than in 2015 [Figure 1A]. Item DIs were relatively constant across the study period [Figure 1B].

Approximately half of the items had an acceptable DIFI (53.4%), while the other half were either difficult (20.8%) or easy (25.9%). The DE was directly related to the DIFI, with DEs of 91.87%, 85.83% and 64.13% for difficult, acceptable and easy items, respectively (*P* <0.005). Items were nearly equally distributed between poor, acceptable and excellent DIs. The DE was 83.24% and 83.33% for items with excellent and acceptable DIs, respectively, compared to 77.56% for items with poor discrimination (*P* <0.005) [Table 3]. There was a significant dome-shaped correlation between DIFI and DI (r = 0.162; *P* = 0.010), with the highest DIs occurring in the acceptable DIFI range and decreasing for DIFIs in the difficult range [Figure 2]. The average $KR_{20}$ coefficient value was 0.76.

**Table 3:** Correlation between difficulty index and discrimination index with distractor efficiency and action proposed of multiple choice questions in end-of-rotation paediatric examinations at the Arabian Gulf University, Manama, Bahrain (N = 800)

| Index | n (%) | DE % | *P* value | Proposed action |
|---|---|---|---|---|
| **DIFI** | | | | |
| Difficult | 166 (20.8) | 91.87 | <0.005* | Review |
| Acceptable | 427 (53.4) | 85.83 | | Store and review |
| Easy | 207 (25.9) | 64.13 | | Discard |
| **DI** | | | | |
| Poor | 254 (31.8) | 77.56 | <0.005† | Discard |
| Acceptable | 267 (33.4) | 83.33 | | Store and review |
| Excellent | 279 (34.9) | 83.24 | | Store |

*DE = distractor efficiency; DIFI = difficulty index; DI = discrimination index.*

**The DE was significantly different for difficult, acceptable and easy items.*
*†Poor DIs had a significantly lower DE than both acceptable and excellent DIs.*

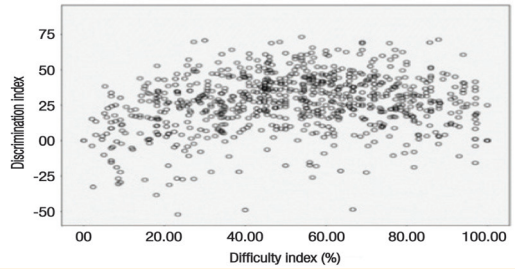


**Figure 2:** Scatter plot showing the relationship between difficulty index and discrimination index among multiple choice question items in end-of-rotation paediatric examinations at the Arabian Gulf University, Manama, Bahrain (N = 800).

## Discussion

In the current study, out of 16 summative examinations and 800 items, the mean DIFI of individual tests was acceptable. Items with a high DIFI mostly occurred in examination papers from 2015 and 2016, while items with a low DIFI mostly occurred in 2013 examination papers. It is likely that this finding reflects recent improvements in MCQ construction by the AGU Examination Committee. The DIFI results of the current study were comparable to those of other institutions, although relative incentives and test conditions are unlikely to have been the same. Mitra *et al.* reported mean DIFIs ranging from 64–89% among 12 summative assessments in their foundation programme conducted between 2003 and 2006.[12] Other studies have reported mean DIFIs of 39.4 ± 21.4% and 63.06 ± 18.95%, respectively.[9,13] Keralia *et al.* reported mean DIFIs between 47.17–58.08% in MCQ items from 10 summative papers.[14] Sharif *et al.* reported a mean DIFI of 49 ± 31% in 2,445 MCQs.[15] In the basic medical sciences component of a nursing licensure examination, Lin *et al.* found the DIFI to range from 10–93%, with a mean of 48%.[16]

Karelia *et al.* reported that 61 ± 8.43%, 24 ± 4.08% and 15 ± 7.07% of items in pharmacology summative tests were acceptable, very easy and very difficult, respectively.[14] In the current study, 53.4%, 25.9% and 20.8% of items fell within these same categories. The authors recommend selecting MCQs with lower DIFIs for fundamental topics that students will probably know; moreover, starting the examination with such questions will raise the students' confidence. Similarly, MCQs with a high DIFI should be located nearer the end of the paper in order to discriminate between high- and low-achievers. With regards to DI, a nearly equivalent percentage of items in the current study were in the poor, acceptable and excellent ranges (31.8%, 33.4% and 34.9%, respectively). Lin *et al.* reported that 28.8% of MCQ items in the basic medical sciences section had a DI of <0.2.[16] Other studies have reported mean DIs of 0.14 ± 0.19, 0.356 ± 0.17, 0.19 ± 0.30 and 0.33 ± 0.18.[6,9,13,15] Items with poor DIs usually result in low scores due to the use of incorrect answer keys, confusing stems or areas of controversy.[17,18] Such items should be removed from the question bank as they fail to discriminate between strong and weak academic performances.

Constructing plausible distractors and decreasing NFDs is essential to improve the quality of MCQs.[19] Therefore, items may need to be modified if students constantly avoid choosing certain distractors. In the current study, most questions had less than two NFDs, with a mean DE of 66.5–90.00%. Other studies have reported a mean DE of 88.6 ± 18.6% and 63.97 ± 33.56%.[9,13] Over the study period, there was a gradual improvement in mean DE from 2013 (70.88%) to 2016 (86.88%); this was likely due to the continuous improvement activities of the AGU Examination Committee. This improvement is also reflected in the annual number of NFDs. Items with zero NFDs increased from 32% in 2013 to 44%, 58% and 59.5% in 2014, 2015 and 2016, respectively, while items with three NFDs decreased from 9% in 2013 to 3.5%, 2% and 1% in 2014, 2015 and 2016, respectively. Items with four NFDs decreased from 4% in 2013 to 0%, 0.5% and 0% in 2014, 2015 and 2016, respectively.

Items with high NFDs reduce both the DE and DI, but increase the DIFI; thus, the item will be easy for the students and act as a poor discriminator of academic performance. In the current study, the DE was significantly higher among difficult items compared to acceptable and easy items as well as significantly higher among items with excellent and acceptable DIs over poor ones. Difficult items with excellent DE values need to be reviewed for possible language confusion, sufficient subject coverage or inappropriately chosen material according to the student's level of learning. In contrast, easy items with low DE values should be discarded, while items with acceptable DIFI and DE values can be stored and reviewed for improvement. It is often necessary to revise items in which the distractor is selected more often than the correct answer.[20] The number of NFDs also affects DI, in that items with lower NFDs are associated with acceptable or excellent DIs. The current study found that items with excellent and acceptable DIs had a significantly higher DE than items with a poor DI. The authors recommend discarding items with poor DIs and low DEs, while retaining items with acceptable or excellent DIs and high DEs.

In the current study, items with NFDs of zero, one and two had acceptable DIFIs and DIs, while items with NFDs of three and four had higher DIFIs and poorer DIs. Mukherjee *et al.* reported a similar

association, with DIFIs of 32.5%, 51.36%, 71.11% and 87.08% for items with zero, one, two and three NFDs, respectively, in a community medicine assessment; only items with NFDs of one and two had acceptable DIs (0.396 and 0.404, respectively), while items with NFDs of zero and three had poor DIs (0.023 and 0.195, respectively).[21] Items which reflect fundamental knowledge should be retained each year to determine whether all students continue to answer them correctly. While some may argue that the inclusion of more options in an MCQ reduces the 'guessing effect', others have demonstrated that additional options beyond three do not make much difference; in fact, reducing the list of available responses to three options can actually improve psychometric features.[22,23]

Furthermore, it is easier to develop three rather than four or five MCQ options and more effective to have fewer options with a greater number of functional distractors in comparison to increased options and more NFDs. Tarrent *et al.* suggested including three instead of four options, as such questions require less time to be constructed and the performance for both is equal.[24] A meta-analysis of 80 years of research concluded that three options are optimal for MCQ items, resulting in a reduction in the amount of time required to prepare each MCQ and allowing more questions to be set per examination.[19] In addition, this will increase subject exposure and improve the reliability and validity of the test due to the inclusion of more high-quality items. According to the dome-shaped correlation between DIFI and DI in the current study, items with DIFIs falling in the difficult or easy categories had significantly poorer DIs. Sim *et al.* similarly reported that maximum DI values were seen with DIFIs between 40–74%.[25] The reliability coefficient in the current study was 0.76, which is less than excellent but still within the desirable range.[2,10,11]

Constructing high-quality MCQs is essential to accurately assess student performance. Overall, for students who know the material covered by the examination, NFDs add little to the performance of a test item; in contrast, increasing the number of distractors decreases the likelihood of students accidentally choosing the correct answer by guesswork. An item analysis of questions is recommended for all examinations in order to continuously update the question bank by keeping items with acceptable indices and revising or discarding others. In the authors' experience, it is usually better to construct an examination with the input of an examination committee in order to improve the quality of the questions. Special training programmes or workshops should be offered to the members of such committees in order to hone their skills in preparing effective MCQs. Further research at AGU is recommended to determine any future improvements in MCQ preparation. Conducting similar studies for examinations in other disciplines at AGU would also be useful.

## Conclusion

Item analyses can be valuable to strengthen an MCQ bank in order to ensure the items have an acceptable DIFI, acceptable or excellent DI and high DE. The item analysis of paediatric end-of-rotation examinations at AGU indicated that a considerable percentage of test items had acceptable mean DIFIs and DIs. However, some items needed to be discarded or revised. Using three or four rather than five options in an MCQ is recommended.

### ACKNOWLEDGEMENTS

The researchers would like to thank the Assessment Office at AGU for providing the MCQ database and helping with the item analysis.

### CONFLICT OF INTEREST

The authors declare no conflicts of interest.

### FUNDING

No funding was received for this study.

## References

1. Case SM, Swanson DB. Extended-matching items: A practical alternative to free-response questions. Tech Learn Med 1993; 5:107–15. doi: 10.1080/10401339309539601.
2. Cronbach LJ, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. Educ Psycol Meas 2004; 64:391–418. doi: 10.1177/0013164404266386.
3. Bloom BS, Hastings JT, Madaus GF. Handbook on Formative and Summative Evaluation of Student Learning. New York, USA: McGraw-Hill, 1971. P. 103.
4. Skakun EN, Nanson EM, Kling S, Taylor WC. A preliminary investigation of three types of multiple choice questions. Med Educ 1979; 13:91–6. doi: 10.1111/j.1365-2923.1979.tb00928.x.
5. Skakun EN, Nanson EM, Taylor WC, Kling S. An investigation of three types of multiple choice questions. Annu Conf Res Med Educ 1977; 16:111–16.
6. 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.
7. 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. doi: 10.1186/1472-6920-9-40.
8. Kelley TL. The selection of upper and lower groups for validation of test items. J Educ Psychol 1939; 30:17–24. doi: 10.1037/h0057123.
9. 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.

10. Bland JM, Altman DG. Cronbach's alpha. BMJ 1997; 314:572. doi: 10.1136/bmj.314.7080.572.

11. Nunnally JC, Bernstein IH. Psychometric Theory, 3rd ed. New York, USA: McGraw-Hill, 1994.

12. 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. Int EJ Sci Med Educ 2009; 3:2–7.

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 Ahemdabad, Gujarat. Indian J Community Med 2014; 39:17–20. doi: 10.4103/0970-0218.126347.

14. 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 M.B.B.S students. Int EJ Sci Med Educ 2013; 7:41–6.

15. 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.

16. Lin LC, Tseng HM, Wu SC. Item analysis of the registered nurse license exam by nursing candidates from vocational nursing high schools in Taiwan. Proc Natl Sci Counc Repub China D 1999; 9:24–31.

17. Kaur M, Singla S, Mahajan R. Item analysis of in use multiple choice questions in pharmacology. Int J Appl Basic Med Res 2016; 6:170–3. doi: 10.4103/2229-516X.186965.

18. Mackenzie J. Vague and ambiguous questions on multiple-choice exercises: The case for. Educ Philos Theory 1994; 26:23–33. doi: 10.1111/j.1469-5812.1994.tb00198.x.

19. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educ Meas Issues Pract 2005; 24:3–13. doi: 10.1111/j.1745-3992.2005.00006.x.

20. Tomak L, Bek Y. Item analysis and evaluation in the examinations in the Faculty of Medicine at Ondokuz Mayis University. Niger J Clin Pract 2015; 18:387–94. doi: 10.4103/1119-3077.151720.

21. Mukherjee P, Lahiri SK. Analysis of multiple choice questions (MCQs): Item and test statistics from an assessment in a medical college of Kolkata, West Bengal. IOSR J Dent Med Sci 2015; 14:47–52. doi: 10.9790/0853-141264752.

22. Nwadinigwe PI, Naibi L. The number of options in a multiple-choice test item and the psychometric characteristics. J Educ Pract 2013; 4:189–96.

23. Vegada B, Shukla A, Khilnani A, Charan J, Desai C. Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. Indian J Pharmacol 2016; 48:571–5. doi: 10.4103/0253-7613.190757.

24. Tarrant M, Ware J. A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. Nurse Educ Today 2010; 30:539–43. doi: 10.1016/j.nedt.2009.11.002.

25. 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.