



# Item Analysis of Multiple Choice and Extended Matching Questions in the Final MBBS Medicine and Therapeutics Examination

Alok Kumar <sup>1,\*</sup>, Colette George<sup>1</sup>, Michael Harry Campbell <sup>1</sup>, Kandamaran Krishnamurthy <sup>1</sup>, Paula Michele Lashley<sup>1</sup>, Virendra Singh<sup>2</sup>, Shastri Motilal<sup>2</sup>, Sateesh Sakhamuri<sup>2</sup>, Tamara Thompson<sup>3</sup>, Corrine SinQuee-Brown<sup>4</sup>, Bidyadhar Sa<sup>2</sup> and Md Anwarul Azim Majumder <sup>1</sup>

<sup>1</sup>The university of the West Indies, Cave Hill, Barbados

<sup>2</sup>The University of the West Indies, St. Augustine, Trinidad and Tobago

<sup>3</sup>The University of the West Indies, Mona, Jamaica

<sup>4</sup>The University of the West Indies, Nassau, Bahamas

\*Corresponding author: The University of the West Indies, Cave Hill, Barbados. Email: [alokkumar.uwichill@gmail.com](mailto:alokkumar.uwichill@gmail.com)

Received 2022 June 29; Revised 2022 August 10; Accepted 2022 September 11.

## Abstract

**Background:** Most universities around the world use the multiple-choice question (MCQ) examination format to evaluate medical education. However, the suitability and advantages of traditional MCQs and extended matching questions (EMQs) continue to be debated.

**Objectives:** This study mainly aimed to perform a comprehensive comparative analysis of the performance of the EMQ and traditional MCQ formats in the final MBBS exit examination.

**Methods:** We conducted an item analysis of 80 EMQs, and 200 MCQs administered to 532 examinees across the four campuses of the University of the West Indies during the final MBBS medicine and therapeutics examination of 2019. Exam performance measures included central tendency, item discrimination, reliability, item difficulty, and distractor efficacy.

**Results:** For the 532 students who sat the exam, the highest, lowest, and mean ( $\pm$  SD) scores for the EMQs were 93, 41, and 62.7 ( $\pm$  7.4), respectively; for the MCQs, the respective values were 82, 41, and 62.7 ( $\pm$  7.4). The predictive value of the EMQ and MCQ grades individually in the overall failure was 0.67 (95% CI = 0.39, 0.87) and 0.89 (95% CI = 0.65, 0.98), respectively. KR-20 coefficients for the EMQs and MCQs ranged from 0.52 to 0.70 and 0.71 to 0.79, respectively. The proportion of questions with two or more functional distractors was consistently higher for the MCQs than for the EMQs in all four cohorts of students.

**Conclusions:** The MCQs were more predictive of the overall failure and had higher inter-item reliability, making the MCQ format more suitable for high-stakes examinations.

**Keywords:** MBBS Examination, Item Analysis, Multiple Choice Questions, Extended Matching Questions

## 1. Background

Most universities around the world use the multiple-choice question (MCQ) examination format in medical education to facilitate testing in large classes and provide better reliability, validity, and objectivity compared to other formats for evaluation (1-4). MCQs are especially suitable for summative exit and licensing examinations in medical sciences (5, 6). In the traditional MCQ, which is the most widely used format for assessment in medical sciences, students are required to select the single best answer from a short list of 4 or 5 choices. However, it has been argued that the traditional MCQ format does not adequately assess the higher levels in the cognitive domain of Bloom's taxonomy (7).

Extended matching questions (EMQs) are multiple-choice items tapping a particular theme of interest organized into sets that use one list of options for all items in the set. A well-constructed EMQ set includes four components: a theme; an option list; a lead-in statement; and at least two item stems (7). Extended matching questions provide a good alternative to MCQs (8, 9). Extended matching questions can be used to evaluate clinical scenarios, provided that examiners construct sufficiently long option lists to minimize cueing and adequately assess clinical reasoning (10-12). Extended matching questions may offer advantages relative to MCQs in both basic and clinical examinations through minimizing cueing effects (11, 12).

Although MCQs may be a standard assessment modality for the cognitive domain of medical education, the suit-

ability and advantages of the different types of MCQs continue to be debated. Recent studies have shown that traditional MCQs may be superior to EMQs in discriminating poor performing students (13, 14).

The University of the West Indies (UWI), a regional university with campuses and medical faculties in Barbados, Jamaica, Trinidad, and Bahamas, has an annual enrollment of approximately 650 medical students in the 5-year MBBS degree. On completion of the final (fifth) year of the MBBS, students must sit a final exit examination in the three major disciplines of medicine and therapeutics, obstetrics and gynecology, and surgery. Students passing this examination are eligible to be provisionally licensed as medical practitioners in most English-speaking Caribbean countries. Each of these three examinations has a written component for assessment in the cognitive domain and a clinical component in the form of objective structured clinical examination (OSCE) for assessment in the affective and sensory domains (15).

The written component of the medicine and therapeutics exit examination comprises a combination of EMQs and MCQs (16). In an ongoing effort to improve quality, we revisited the effectiveness of the question format for this examination. A large number of examinees from four geographically diverse campuses taking the same written examination in their final year of study provided an excellent opportunity for robust assessment of the performance of items in this examination.

## 2. Objectives

The present study's main objective was to conduct a comprehensive comparative analysis of the EMQ and MCQ formats of the written medicine and therapeutics component of the final MBBS examination.

## 3. Methods

The data for this study were collected from the written medicine and therapeutics component of the final MBBS examination of 2019 at UWI. We conducted an item analysis of 80 EMQs, and 200 MCQs administered to 532 examinees across the four UWI campuses. Specifically, the written exam consisted of two papers, each with two sections (A and B). Section A had 40 thematic EMQs, and section B had 100 5-choice single-best-answer MCQs. The same question papers were used on all four campuses. A university examiner (UE) selected all questions from a question bank using an established blueprint to ensure a representative

distribution of content. Two independent external examiners reviewed and approved the finalized exam papers. Every year, faculty members who have participated in workshops on writing effective exam items write new questions and submit them to the UE. Submitted questions are peer-reviewed and standard set for the level of item difficulty using the Modified Angoff method (17). These newly vetted items are continuously added to the question bank maintained by UWI.

The examination was administered simultaneously on all four campuses in proctored examination centers using paper and pencil. The answer sheets for the candidates on all four campuses were collected by the UE and marked using the Scantron® optical scanner (18). Scantron Assessment Solutions generated a database of scores for each candidate and provided item analysis for each section in both papers. Further analysis was completed using SPSS® v25, 2017 (IBM Corporation). The data were anonymized by removing student identification numbers and assigning a duplicate ID used to link the exam scores to the candidates without disclosing their identity.

This study involved analysis of de-identified examination data and, therefore, was exempt from review by the research ethics committee. The authors followed the Declaration of Helsinki during all phases of the study.

Exam performance measures included central tendency, item discrimination, reliability, item difficulty, and distractor efficacy. We computed point-biserial discrimination index (DI) scores for all items and used a threshold of 0.2 to establish adequate discriminability (19). We calculated Kuder-Richardson formula 20 (KR-20) to assess the internal consistency of the MCQ and EMQ sections (19). Item difficulty index (p) scores for each item in both sections were also analyzed (20). Items with P-value < 0.3 or > 0.8 were considered non-discriminatory. Further, we calculated distractor efficiency (DE) scores for the incorrect options on each question. Distractors selected by > 5% of the students were considered to be non-functional. Distractor efficiency was acceptable if the items had two or more functional distractors (DE > 50%). We used the overall exam failure as the criterion for calculating the predictive value of the EMQ and MCQ components. Candidates are required to pass all exam components; therefore, students who failed one or more of the six different components of this examination failed the overall MBBS final examination. Differences in MCQ and EMQ scores were assessed using one-way analysis of variance (ANOVA). Also, 50% was the minimum pass score; students scoring 65% earned honors; and those scoring 75% achieved distinction.

## 4. Results

Five hundred and thirty-two (532) students took the written medicine and therapeutics component of the final MBBS exam, of whom 63.6% were females, and 36.4% were males. The students were divided into four cohorts (arbitrarily numbered 1 to 4 to avoid identifiable comparisons) representing the different medical campuses of UWI, with 260, 194, 43, and 35 students in the cohorts 1, 2, 3, and 4, respectively. Overall, 495 (93.1%; 95% CI = 90.5%, 95%) students were taking this exam for the first time, and 37 (7%; 95% CI = 5%, 9.5%) students were taking the exam for the second or third time. Of the 532 students who sat the exam, 513 passed, and 19 failed.

### 4.1. Scoring Pattern for EMQs and MCQs

Comparisons of scores from the EMQ and MCQ sections are shown in [Table 1](#). The maximum achievable score was 100 for each section. For the 532 students who sat the exam, the highest, lowest, and mean ( $\pm$  SD) scores for EMQs were 93, 41, and 69.0 ( $\pm$  9.8), respectively; for MCQs, the respective values were 82, 41, and 62.7 ( $\pm$  7.4). The difference between scores from the EMQ and MCQ sections for all 532 students was statistically significant ( $P < 0.0001$ ). Based on the EMQ scores alone, 14 (2.6%; 95% CI = 1.5%, 4.5%) students did not achieve passing scores; 261 (49.1%; 95% CI = 44.7%, 53.4%) students had passing scores; and 257 (48.3%; 95% CI = 44.0%, 52.7%) students performed at the honors level. The corresponding figures from the MCQ section were 24 (4.5%; 95% CI = 3.0%, 6.7%) students; 402 (75.6%; 95% CI = 71.6%, 79.1%) students; and 106 (19.9%; 95% CI = 16.7%, 23.6%) students. The proportion of students failing in the EMQ section was not significantly different from that in the MCQ section (OR = 0.57; 95% CI = 0.29, 1.12;  $P = 0.099$ ). The positive predictive value of the EMQ scores for overall failure in the written component of the medicine and therapeutics exam was 0.67 (95% CI = 0.39, 0.87) with likelihood ratios (conventional) of 54.0 (95% CI = 20.4, 142.6). The positive predictive value of the MCQ scores for overall failure was superior: 0.89 (95% CI = 0.65, 0.98) with likelihood ratios of 188.11 (95% CI = 46.21, 766.12).

### 4.2. Discrimination Index (DI or $r$ ) Values for EMQs and MCQs

The mean DI scores for the EMQ and MCQ components of the examination are shown in [Table 2](#). There were no statistically significant differences in DI scores by question type between the four cohorts. The proportion of EMQs and MCQs with a DI  $> 0.2$  (acceptable level of discrimination) is shown in [Figure 1](#). The proportion of questions with a DI  $> 0.2$  was higher for the EMQs compared to the MCQs in all four cohorts of students, although insignificant. OR

for the proportion of EMQs that were acceptably discriminatory when compared with the MCQs was 1.62 (95% CI = 0.87, 3.02;  $P = 0.13$ ), 1.19 (95% CI = 0.69, 2.07;  $P = 0.05$ ), 1.54 (95% CI = 0.90, 2.62;  $P = 0.11$ ), and 1.12 (95% CI = 0.66, 1.92;  $P = 0.67$ ) for the cohorts 1, 2, 3, and 4, respectively.

### 4.3. Reliability (Internal Consistency) of EMQs and MCQs

Internal consistency values for EMQs and MCQs are summarized by campus cohort in [Table 3](#). KR-20 coefficients for EMQs ranged from 0.66 to 0.70 and 0.52 to 0.69 in papers 1 and 2, respectively. The corresponding values for MCQs were in the range of 0.75 - 0.79 and 0.71 - 0.77 for papers 1 and 2, respectively.

### 4.4. Difficulty Index for EMQs and MCQs

Difficulty Index (DIFI) that refers to the proportion of students correctly answering a question, for all the questions used in this examination, is shown in [Figure 2](#). The proportion of EMQs with a DIFI value between 0.3 and 0.8 ranged between 47.5% - 75% for the four cohorts of students. The corresponding figure for the MCQs ranged from 43% to 62.5%. The difference in the proportion of questions in the three categories of the p-value ( $< 0.3$ , 0.3 - 0.8, and  $> 0.8$ ) was statistically significant for the cohort 1 ( $P \leq 0.0001$ ), but insignificant for any other cohort.

### 4.5. Distractor Efficiency of EMQs and MCQs

The proportions of EMQs and MCQs with functional distractors for each cohort are shown in [Figure 3](#). The proportion of questions with two or more FDs was consistently higher for the MCQs compared to the EMQs for all cohorts. The proportion of questions with two or more FDs ranged from 42.5% to 53.5% and 50.5% to 69% across the cohorts for the EMQ and MCQ items, respectively. However, the difference was statistically significant only for the cohort 2 (OR = 0.49; 95% CI = 0.29, 0.83;  $P = 0.007$ ) and the cohort 4 (OR = 0.52; 95% CI = 0.31, 0.89;  $P = 0.02$ ).

## 5. Discussion

Examinations required for medical qualification and certification of fitness to practice must be designed with careful attention to key issues, including blueprinting, validity, reliability, and standard setting, as well as clarity about their formative or summative function (12). Items used in assessment should be sufficiently discriminatory for minimally competent and high-achieving students and reasonably easy to construct. Additionally, an assessment should reflect key educational objectives in all components of the cognitive domain of Bloom's taxonomy (21).

**Table 1.** Comparison of the Extended Matching Question and Multiple-choice Question Scores Across the Four Campuses and Overall Using ANOVA in the Final Exit MBBS Medicine and Therapeutics Examination of the University of the West Indies (2019)

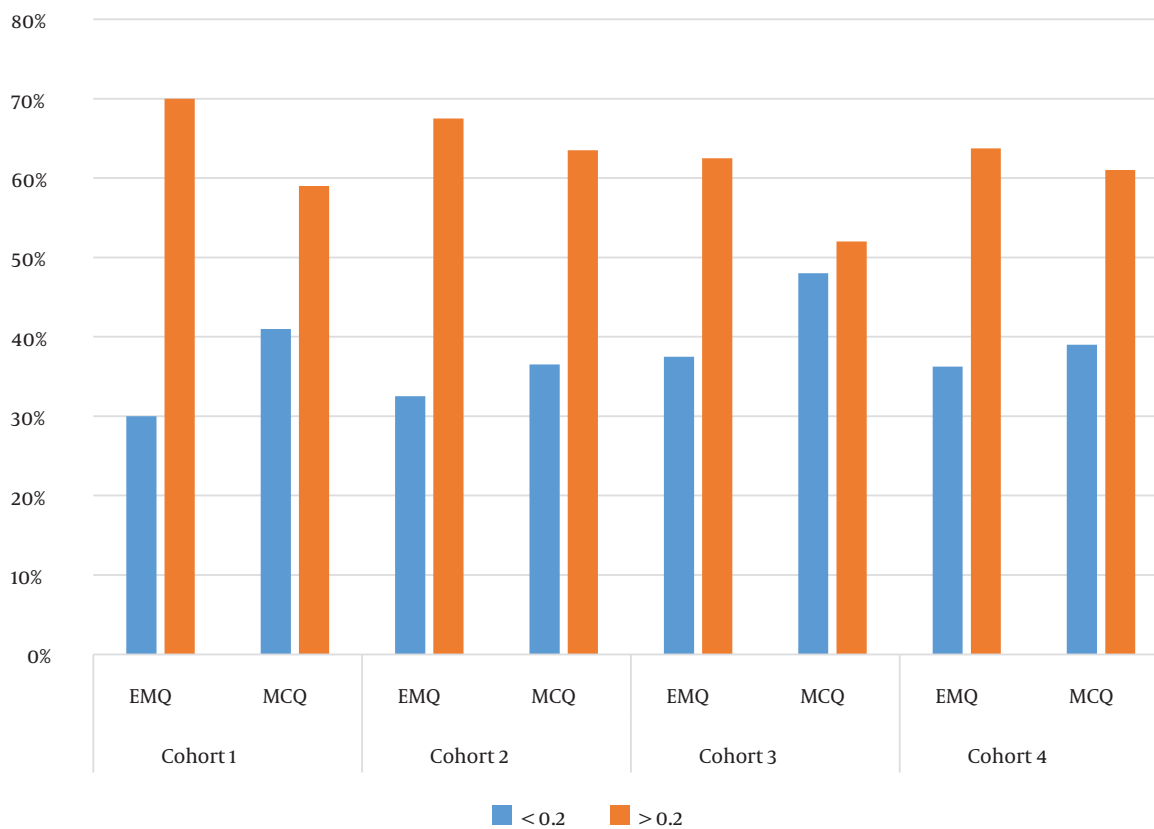
Variables	Cohort 1		Cohort 2		Cohort 3		Cohort 4		Overall	
	EMQ	MCQ	EMQ	MCQ	EMQ	MCQ	EMQ	MCQ	EMQ	MCQ
Highest score	91	77	93	81	93	82	84	72	93	82
Lowest score	43	44	44	41	51	51	41	44	41	41
Mean score	69.3	63.2	68.5	61.8	71.4	66.4	66.3	59.2	69.0	62.7
Standard deviation	9.9	6.7	9.7	7.7	9.5	7.9	9.5	8.0	9.8	7.4
Standard error	0.616	0.4	0.7	0.6	1.4	1.2	1.6	1.3	0.4	0.3
Variance	98.65	44.9	93.3	59.4	90.2	62.1	89.6	63.3	96.0	54.8
P-value	< 0.0001		< 0.0001		0.0088		0.0012		< 0.0001	

Abbreviations: EMQs, extended matching questions; DI, discrimination index; MCQs, multiple-choice questions.

**Table 2.** Comparison of the Mean Discrimination Index Scores of the Extended Matching Questions and Multiple-choice Questions Used in the Final Exit Medicine and Therapeutics Examination of the University of the West Indies (2019) Using Paired t-Test

Cohort	Median DI EMQ	Mean $\pm$ SD DI EMQ	Median DI MCQ	Mean $\pm$ SD DI MCQ	P-Value *(t-Test)
Cohort 1	0.25	0.37 $\pm$ 0.25	0.24	0.27 $\pm$ 0.47	0.0680
Cohort 2	0.31	0.36 $\pm$ 0.23	0.28	0.27 $\pm$ 0.55	0.0783
Cohort 3	0.31	0.34 $\pm$ 0.35	0.21	0.23 $\pm$ 0.61	0.0526
Cohort 4	0.29	0.33 $\pm$ 0.32	0.29	0.24 $\pm$ 0.56	0.0914

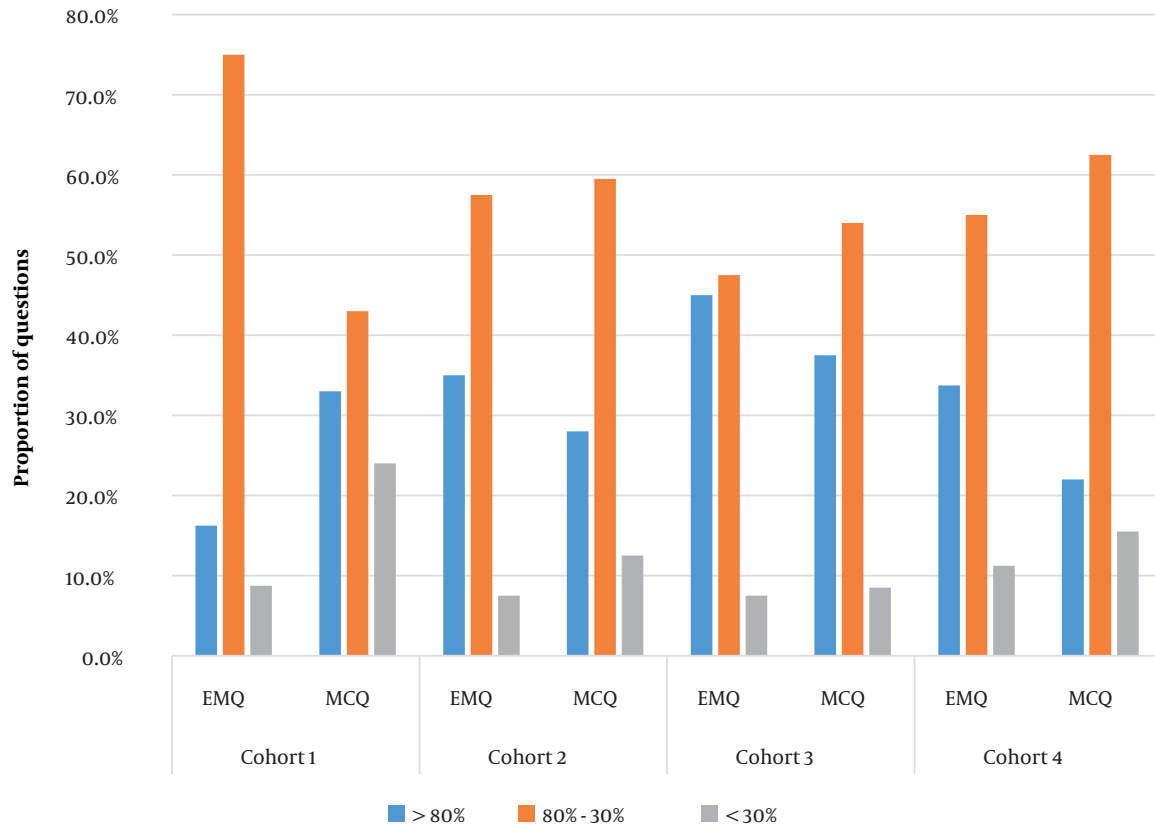
Abbreviations: EMQs, extended matching questions; DI, discrimination index; MCQs, multiple-choice questions.

**Figure 1.** The discrimination index scores of the extended matching questions (EMQs) and the multiple-choice questions (MCQs) in the four cohorts of students sitting the final exit MBBS medicine and therapeutics examination of the University of the West Indies (2019).

**Table 3.** Comparisons of the KR-20 Values for the Extended Matching Questions and the Multiple-choice Questions Used in the Final Exit MBBS Medicine and Therapeutics Examination of the University of the West Indies (2019)

Cohort	Paper 1		Paper 2	
	EMQs	MCQs	EMQs	MCQs
Cohort 1	0.69	0.72	0.69	0.71
Cohort 2	0.66	0.79	0.66	0.73
Cohort 3	0.70	0.75	0.61	0.77
Cohort 4	0.68	0.78	0.52	0.72

Abbreviations: EMQs, extended matching questions; MCQs, multiple-choice questions.

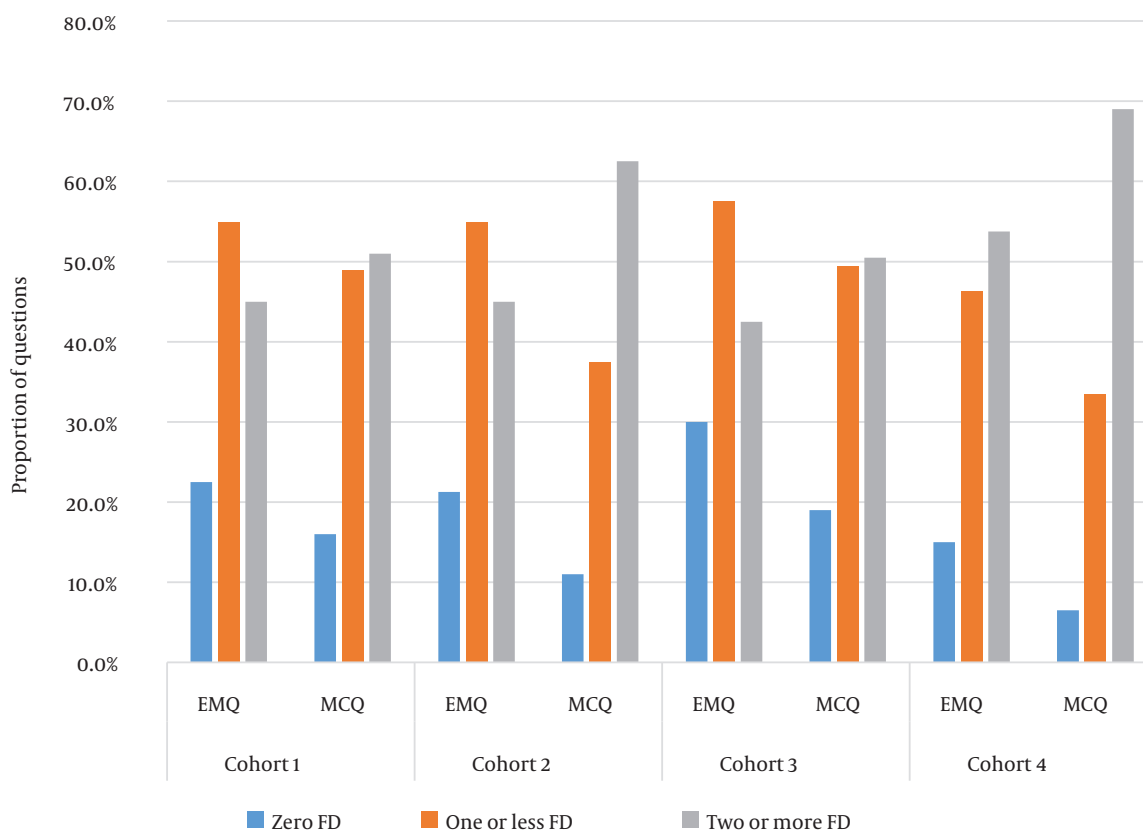
**Figure 2.** The difficulty index (percentage of students correctly answering a question) scores for the extended matching questions (EMQs) and the multiple-choice questions (MCQs) in the four cohorts of students sitting the final exit MBBS medicine and therapeutics examination of the University of the West Indies, (2019).

High reliability is especially important for the final MBBS examination, given its function to license medical practitioners (22). Assessment processes should be continuously evaluated, and the feedback should be used to improve subsequent examinations. This study compared the reliability, discrimination index, and quality of EMQs and MCQs constructed by faculty members trained in item writing and standard set using the modified Angoff method (17) for the final MBBS examination completed by students from campuses in four member countries of the same regional university with the same curriculum and learning objec-

tives. These attributes make this study unique and, to our knowledge, the first such study to be reported in the medical education literature.

### 5.1. Scoring Pattern for EMQs and MCQs

In the current study, the overall mean score (Table 1) for the EMQs ( $69\% \pm 9.8\%$ ) was significantly higher than that for the MCQs ( $62.7\% \pm 7.4\%$ ). Significantly higher mean scores for the EMQs were seen for all four cohorts of students who attempted this examination. Similar findings



**Figure 3.** The distractor analysis of the extended matching questions (EMQs) and the multiple-choice questions (MCQs) in the four cohorts of students sitting the final exit MBBS Medicine and Therapeutics Examination of the University of the West Indies, (2019).

have been reported in another comparative study of different modalities of assessment used in the MBBS examination (23). The scores from the EMQs had a larger spread with higher standard deviation values and variance, which would be advantageous and more discriminatory for feedback to students and teachers in the formative assessment. An important finding from this study was that the score from the MCQs had a higher positive predictive value for the overall failure in the written examination when compared to the score from the EMQs. One criticism of EMQs in medical assessment has been that they are less capable of detecting poor performers compared to MCQs (13). Our study provides strength to this criticism.

### 5.2. Discrimination Index (DI or $r$ ) for EMQs and MCQs

The mean DI (24) was higher for the EMQs than for the MCQs in all four cohorts, although the difference was not statistically significant (Table 2). The mean DIs for the EMQs (range:  $0.33 \pm 0.32$  -  $0.37 \pm 0.25$  among the four cohorts) and the MCQs (range:  $0.23 \pm 0.61$  -  $0.27 \pm 0.47$ ) were com-

parable to DIs for MCQs in previous studies (25, 26). Additionally, the proportion of questions with a DI > 0.02 was higher for the EMQs than for the MCQs in all four cohorts, although insignificant. As a general rule, items with DI values < 0.20 are considered poor, indicating that they should be eliminated or revised, and items with DI values > 0.20 are considered fair to good (27). In the present analysis, between 50% - 70% of both EMQs and MCQs had DI values > 0.20, which are comparable to those reported for similar high-stakes examinations (28, 29). The high proportion of EMQs and MCQs with fair to good DIs in this exam analysis supports the validity of the written assessment tool in this examination (27).

### 5.3. Reliability (Internal Consistency) for EMQs and MCQs

The KR-20 for the EMQs, ranging from 0.52 to 0.70, was lower than that for the MCQs, which ranged between 0.71 and 0.79 (Table 3). The KR-20 index ranges from 0 to 1, and it is a measure of inter-item reliability. A higher value for an exam indicates a stronger relationship between items on

the test. A low reliability coefficient may be reflected when a test covers multiple topics and also reflects the total number of test questions. Generally, for a high-stakes or licensing examination, a KR-20 value closer to 0.80 is preferred. Of note, there were 80 EMQs and 200 MCQs in this examination. The lower KR-20 for the EMQs may partly be due to the lower number of EMQs used in this examination. Also, this examination covered a number of specialty topics for which a KR-20 value of 0.50 would be an acceptable lower limit (19).

#### 5.4. Difficulty Index (DIFI or *p*) for EMQs and MCQs

Analysis of difficulty revealed that the proportion of questions in each of the three categories based on *p*-values ( $< 0.3$ ,  $0.3 - 0.8$  and  $> 0.8$ ) was not significantly different between the EMQs and the MCQs in three of the four cohorts of students (Figure 2). Questions with *p*-values  $< 0.3$  and  $> 0.8$  are usually non-discriminatory. In the present study, the proportion of questions in each cohort with a *P*-value between 0.3 and 0.8 ranged from 47.5% to 75% and from 43% to 62.5% for the EMQs and the MCQs, respectively. Overall, 55.25% and 49.43% of the questions had a DI between 0.3 and 0.8 for the EMQs and the MCQs, respectively. However, the difference was not statistically significant. The *p*-values for the EMQs and the MCQs in our study compared well with those from other studies of medical examinations (28, 30-32).

#### 5.5. Distractor Efficiency of EMQs and MCQs

Functional distractors for MCQs decrease correct guessing and cueing. In fact, one advantage of EMQs over MCQs is an increase in the number of distractors, decreasing the likelihood of guessing and cueing (33, 34). In the present study, a higher proportion of the MCQs had two or more functional distractors when compared to the EMQs in all four cohorts of students, although the difference was statistically significant for only one of the four cohorts. The increased number of distractors in EMQs makes item writing more difficult by requiring more plausible distractors compared to MCQs. It is important to note that having a higher proportion of functional distractors is especially important in EMQs to avoid testing time increase without having its given advantages. Overall, items with two or more functional distractors in both EMQs and MCQs were comparable to those reported from other studies (28, 31, 32). However, of concern was the finding that up to 30% of the EMQs and 19% of the MCQs had no functional distractors. This finding may reflect poor item construction by some examiners, as shown in other studies (35). Repeated use of questions from a bank for successive examinations

may negatively impact the performance of distractors. Although fewer than 15% of items were questions that were repeated from the recent final MBBS examinations, this proportion may have been higher if all of the past examinations were taken into account. With a higher proportion of repeated questions, the distractors may become less effective, and this may have partly contributed to the finding of the high proportion of questions with no functional distractor in this study.

Regular revision and replenishing are required to sustain the viability of the question bank.

The observed wider spread of scores and higher mean of EMQs compared to MCQs suggest that EMQs are more suitable for feedback in formative assessment. However, the MCQ scores were more predictive of overall exam failure on the written component, suggesting that MCQs are more suitable for high-stakes assessments such as the final MBBS examination.

#### 5.6. Conclusions

Although there was no significant difference between the DIs of the EMQ and MCQ items, the MCQs demonstrated higher internal consistency. However, both EMQs and MCQs demonstrated similar levels of difficulty. Also, the EMQs displayed poorer distractor efficiency than the MCQs, a finding that reflects the inherent difficulty in EMQ item construction.

#### Footnotes

**Authors' Contribution:** A K and M A A M, conceptualization, analysis, draft writing of the manuscript and approval of the final manuscript; C G, V S, S M, P M L, S S, T T and C S B, data collection, critical review of the manuscript and approval of the final manuscript; M H C and K K, conceptualization, critical review of the manuscript and approval of the final manuscript; B S, analysis, critical review of the manuscript and approval of the final manuscript.

**Conflict of Interests:** None declared by the authors.

**Data Reproducibility:** The dataset presented in the study is available on request from the corresponding author during submission or after its publication. The data are not publicly available due to the nature of the data-exam results.

**Funding/Support:** Nothing declared by the authors.

#### References

1. Wilson RB, Case SM. JVME v20n3: Extended Matching Questions: An Alternative to Multiple-choice or Free-response Questions. *JVME*. 1993;3:V20-3.

2. Shumway JM, Harden RM, Association for Medical Education in E. AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. *Med Teach*. 2003;**25**(6):569-84. [PubMed ID: 15369904]. <https://doi.org/10.1080/0142159032000151907>.
3. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. *J Family Community Med*. 2006;**13**(3):125-33. [PubMed ID: 23012132]. [PubMed Central ID: PMC3410060].
4. Schuwirth LW, van der Vleuten CP. ABC of learning and teaching in medicine: Written assessment. *BMJ*. 2003;**326**(7390):643-5. [PubMed ID: 12649242]. [PubMed Central ID: PMC1125542]. <https://doi.org/10.1136/bmj.326.7390.643>.
5. Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;**356**(4):387-96. [PubMed ID: 17251535]. <https://doi.org/10.1056/NEJMr054784>.
6. van der Vleuten C. Validity of final examinations in undergraduate medical training. *BMJ*. 2000;**321**(7270):1217-9. [PubMed ID: 11073517]. [PubMed Central ID: PMC118966]. <https://doi.org/10.1136/bmj.321.7270.1217>.
7. Case SM, Swanson DB. *Constructing Written Test Questions For the Basic and Clinical Sciences*. 3rd ed. Philadelphia, USA: National Board of Medical Examiners; 2002.
8. Case SM, Swanson DB. Extended-matching items: a practical alternative to free-response questions. *Teach Learn Med*. 1993;**5**(2):107-15. <https://doi.org/10.1080/10401339309539601>.
9. Beullens J, Van Damme B, Jaspaert H, Janssen PJ. Are extended-matching multiple-choice items appropriate for a final test in medical education? *Med Teach*. 2002;**24**(4):390-5. [PubMed ID: 12193322]. <https://doi.org/10.1080/0142159021000000843>.
10. Beullens J, Struyf E, Van Damme B. Do extended matching multiple-choice questions measure clinical reasoning? *Med Educ*. 2005;**39**(4):410-7. [PubMed ID: 15813764]. <https://doi.org/10.1111/j.1365-2929.2005.02089.x>.
11. Fenderson BA, Damjanov I, Robeson MR, Veloski J, Rubin E. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. *Hum Pathol*. 1997;**28**(5):526-32. [https://doi.org/10.1016/s0046-8177\(97\)90073-3](https://doi.org/10.1016/s0046-8177(97)90073-3).
12. Tabish SA. Assessment methods in medical education. *Int J Health Sci*. 2008;**2**(2):3-7.
13. Eijsvogels TM, van den Brand TL, Hopman MT. Multiple choice questions are superior to extended matching questions to identify medicine and biomedical sciences students who perform poorly. *Perspect Med Educ*. 2013;**2**(5-6):252-63. [PubMed ID: 24203858]. [PubMed Central ID: PMC3824749]. <https://doi.org/10.1007/s40037-013-0068-x>.
14. Medical Science Devison. *Online Assessment: What is the best question type? How many options should students choose from?* London, UK: University of Oxford; 2022. Available from: <https://www.medsci.ox.ac.uk/divisional-services/support-services-1/learning-technologies/faqs/what-is-the-best-question-type-how-many-options-should-students-choose-from>.
15. Bloom BS, editor. *Taxonomy of Educational Objectives. The cognitive domain*. 1956; New York, USA. 1956.
16. Majumder MAA, Kumar A, Krishnamurthy K, Ojeh N, Adams OP, Sa B. An evaluative study of objective structured clinical examination (OSCE): students and examiners perspectives. *Adv Med Educ Pract*. 2019;**10**:387-97. [PubMed ID: 31239801]. [PubMed Central ID: PMC6556562]. <https://doi.org/10.2147/AMEP.S197275>.
17. Angoff WH. Scales, norms, and equivalent scores. In: Angoff WH, editor. *Educational measurement RL Thorndike*. USA: American Council on Education; 1971.
18. Scantron. *Optical Mark Recognition (OMR Scanners)*. Minnesota, United States: Scantron; [cited 12 January 2021]. Available from: <https://www.scantron.com/scanners-forms/optical-mark-recognition-omr-scanners-opscan-series/>.
19. McGahee TW, Ball J. How to read and really use an item analysis. *Nurse Educ*. 2009;**34**(4):166-71. [PubMed ID: 19574855]. <https://doi.org/10.1097/NNE.0b013e3181aaba94>.
20. McDonald ME. *The nurse educator's guide to assessing learning outcomes*. 4th ed. Massachusetts, USA: Jones and Bartlett Learning; 2017. 428 p.
21. Anderson LW, Krathwohl DR. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. London, UK: Longman; 2001.
22. van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;**39**(3):309-17. [PubMed ID: 15733167]. <https://doi.org/10.1111/j.1365-2929.2005.02094.x>.
23. AlShamlan NA, Al Shammari MA, Darwish MA, Sebiany AM, Sabra AA, Alalmaie SM. Evaluation of Multifaceted Assessment of the Fifth-Year Medical Students in Family Medicine Clerkship, Saudi Arabia Experience. *J Multidiscip Healthc*. 2020;**13**:321-8. [PubMed ID: 32256080]. [PubMed Central ID: PMC7093101]. <https://doi.org/10.2147/JMDH.S241586>.
24. Haladyna TM. *Developing and validating multiple-choice test items*. 3rd ed. Oxfordshire, UK: Routledge; 2004. 320 p. <https://doi.org/10.4324/9780203825945>.
25. Mehta G, Mokhasi VR. Item Analysis of Multiple Choice Questions- An Assessment of the Assessment Tool. *Int J Health Sci Res*. 2014;**4**:197-202.
26. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *J Pak Med Assoc*. 2012;**62**(2):142-7.
27. Chiavaroli N, Familiari M. When Majority Doesn't Rule: The Use of Discrimination Indices to Improve the Quality of MCQs. *Biosci Educ*. 2011;**17**(1):1-7. <https://doi.org/10.3108/beej.17.8>.
28. 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):e68-74. [PubMed ID: 29666684]. [PubMed Central ID: PMC5892816]. <https://doi.org/10.18295/squmj.2018.18.01.011>.
29. Lin L, Tseng H, Wu S. Item Analysis of the Registered Nurse License Exam by Nurse Candidates from Vocational Nursing High Schools in Taiwan. *Proc Natl Sci Coun Repub China Part B Life sci*. 1999;**9**:24-30.
30. N Karelia B. 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 e-J Sci Med Educ*. 2013;**7**(2):41-6. <https://doi.org/10.56026/imu.7.2.41>.
31. Vuma S, Sa B. A descriptive analysis of extended matching questions among third year medical students. *Int J Res Med Sci*. 2017;**5**(5):1913-20. <https://doi.org/10.18203/2320-6012.ijrms20171817>.
32. Vuma S, Sa B. A comparison of clinical-scenario (case cluster) versus stand-alone multiple choice questions in a problem-based learning environment in undergraduate medicine. *J Taibah Univ Med Sci*. 2017;**12**(1):14-26. [PubMed ID: 31435208]. [PubMed Central ID: PMC6694941]. <https://doi.org/10.1016/j.jtumed.2016.08.014>.
33. Swanson DB, Holtzman KZ, Allbee K. Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options. *Acad Med*. 2008;**83**(10 Suppl):S21-4. [PubMed ID: 18820493]. <https://doi.org/10.1097/ACM.0b013e318183e5bb>.
34. Hift RJ. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Med Educ*. 2014;**14**:249. [PubMed ID: 25431359]. [PubMed Central ID: PMC4275935]. <https://doi.org/10.1186/s12909-014-0249-2>.
35. 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. [PubMed ID: 19580681]. [PubMed Central ID: PMC2713226]. <https://doi.org/10.1186/1472-6920-9-40>.