ORIGINAL ARTICLE

Descriptive Analysis of the Psychometric Properties of Extended Matching Questions Conducted among Anaesthesia Residents

Annie Sheeba J*, Ravindaran Charulatha, Hemanth Kumar V R
Department of Anaesthesiology, Mahatma Gandhi Medical College and Research Institute, Pondicherry, India

*Mahatma Gandhi Medical College and Research Institute, Sri Balaji Vidyapeeth Pillayarkuppam, Pondicherry, 607402 India

Tel: +91-9655035791
Email: dr.anniej@yahoo.com

Background: Clinical reasoning is one of the core features of clinical competency. Training and assessing clinical reasoning is vital in post-graduate training. Extended matching questions (EMQs) are effective in assessing problem-solving and clinical reasoning abilities, but not commonly used in Postgraduate training. Covid-19 pandemic, which prevented both patient encounter and regular academic activities, warranted the introduction of innovative Teaching-Learning methods to sustain clinical reasoning skills. Hence, we aimed to introduce EMQs in formative assessment among anesthesia residents and analyzed its psychometric properties.

Methods: The study was conducted at the Department of Anaesthesiology, Mahatma Gandhi Medical College and Research Institute (MGMCRI), Pondicherry, India. Four modules of EMQs as part of a formative assessment was conducted among residents (n=20). A total of 40 clinical vignettes and 60 options were administered. Post-validation of the EMQs was done by item analysis. Test reliability was estimated by the Kuder-Richardson 20 formula. Difficulty index (DIF-I), discrimination index (DIS-I), and distractor functionality were analyzed.

Results: The KR-20 reliability coefficient was 0.72. The mean DIF-I was 0.43±0.17, from which 72.5% (29) were in the acceptable range, 20% (8) difficult, and 7.5% (3) easy. The mean DIS-I was 0.28±0.24, where 40% (16) had acceptable, 27.5% (11) excellent, and 20% had poor discrimination. Ninety percent of distractors were functional. The DIS-I exhibited a positive correlation with DIF-I (r= 0.2155, P=0.0185).

Conclusions: The results of the present study indicated that EMQs have acceptable test reliability. The majority of the items (80%) followed the principles of MCQs. We concluded that EMQs can be effectively used as part of the postgraduate assessment to test higher-order knowledge and clinical competency.

Keywords: Extended matching questions, Item Analysis, Difficulty Index, Clinical competency

التحليل الوصفي للخصائص السيكومترية لأسئلة التكيف الموسعة في مساعدي التخدير

الخلفية والهدف: التفكير السريري هو أحد السمات الرئيسية للكفاءة السريرية. يعد تدريس المنطق السريري وتقييمه أمرًا بالغ الأهمية في التعليم بعد التخرج. تعتبر أسئلة المطابقة المكثفة (EMQs) فعالة في تقييم القدرة على حل المشكلات والتفكير السريري ، ولكنها ليست شائعة الاستخدام في دورات الدراسات العليا. يضمن جائحةCovid-19 ، الذي يمنع تعرض المريض للأنشطة الأكاديمية المنتظمة ، إدخال طرق تعليم وتعلم مبتكرة للحفاظ على مهارات التفكير الإكلينيكي. ومن ثم ، فإن هدفنا هو تقديم EMQ في التقييم التكويني بين مساعدي التخدير وتحليل خصائصه السيكومترية.

الطريقة: تم إجراء هذه الدراسة في قسم التخدير وكلية الطب ومعهد أبحاث المهاتما غاندي (MGMCRI) في بونديشيري ، الهند. تم إجراء أربع وحدات EMQكجزء من التقييم البناء بين السكان (٢٠ شخصًا). تم استخدام ما مجموعه ٤٠ تعليقًا توضيحيًا سريريًا و ٦٠ خيارًا. بعد التحقق من الصحة ، تم إجراء EMQsمع تحليل الحالة. تم تقدير موثوقية الاختبار بواسطة صيغة Kuder-Richardson 20. تم تحليل مؤشر الصعوبة (DIF-I) و مؤشر التمييز (DIS-I)و أداء التشتت.

النتائج: كان معامل الموثوقية لـ KR-20 0.72. كان متوسط DIF-I 0.43±0.17 ، منها ٧٢,٥٪ (٢٩) كانت ضمن النطاق المقبول ، و ٢٠٪ (٨) كانت صعبة و ٧,٥٪ (٣) كانت سهلة. كان متوسط DIS-I 0.28±0.24 ، منها ٤٠٪ (١٦) كانت مقبولة ، ٢٧,٥٪ (١١) كانت ممتازة ، و ٢٠٪ كان لديها تمييز ضعيف. تسعون بالمائة من المشتتات كانت وظيفية. أظهر DIS-I ارتباطًا إيجابيًا مع DIF-I (r= 0.2155, P=0.0185).

الخلاصة: تظهر نتائج الدراسة الحالية أن EMQ لديها موثوقية اختبار مقبولة. معظم (٨٠٪) يتبعون مبادئ MCQ. نستنتج أنه يمكن استخدام EMQ بشكل فعال كجزء من تقييم الدراسات العليا لاختبار مستوى أعلى من المعرفة والكفاءة السريرية.

الكلمات المفتاحية: أسئلة المطابقة الواسعة ، تحليل الحالة ، مؤشر الصعوبة ، الكفاءة السريرية

تحلیل توصیفی خصوصیات روان سنجی سوالات تطبیق گسترده در دستیاران بیهوشی

زمینه و هدف: استدلال بالینی یکی از اصلی ترین ویژگی های صلاحیت بالینی است. آموزش و ارزیابی استدلال بالینی در آموزشهای تحصیلات تکمیلی حائز اهمیت است. سؤالات تطبیق گسترده (EMQ) در ارزیابی توانایی حل مسئله و استدلال بالینی مؤثر است، اما معمولاً در دوره های تحصیلات تکمیلی استفاده نمی شود. بیماری همه گیر کووید ۱۹ که مانع از مواجهه با بیمار و فعالیتهای منظم دانشگاهی می شود، معرفی روش های نوآورانه تدریس-یادگیری را برای حفظ مهارت استدلال بالینی تضمین می کند. از این رو، هدف ما معرفی EMQ در ارزیابی تکوینی در بین دستیاران بیهوشی و تجزیه و تحلیل خصوصیات روان سنجی آن است.

روش: این مطالعه در گروه بیهوشی، کالج پزشکی و مؤسسه تحقیقات ماهاتما گاندی (MGMCRI) در شهر پوندیچری هند انجام شد. چهار ماژول EMQ به عنوان بخشی از ارزیابی سازنده در بین دستیاران انجام شد (۲۰ نفر). در مجموع ۴۰ عکس شرح دار بالینی و ۶۰ گزینه استفاده شد. پس از اعتبار سنجی EMQ ها با تجزیه و تحلیل مورد انجام شد. پایایی آزمون با فرمول کودر-ریچاردسون ۲۰ برآورد شد. شاخص دشواری (DIF-I) ، شاخص تبعیض (DIS-I) و عملکرد پراکندگی مورد تجزیه و تحلیل قرار گرفت.

یافته ها: ضریب اطمینان KR-20 72/0 بود. میانگین DIF-I 17/0 43/0 بود که ۷۲,۵٪ (۲۹) در محدوده قابل قبول ، ۲۰٪ (۸) دشوار و ۷,۵٪ (۳) آسان بودند. میانگین DIS-I 24/0 28/0 بود که ۴۰٪ (۱۶) قابل قبول ، ۲۷/۵٪ (۱۱) عالی و ۲۰٪ تبعیض ضعیف داشتند. نود درصد عوامل حواس پرتی عملکردی داشتند. DIS-I همبستگی مثبتی با DIF-I نشان داد (r = 0.2155 ،P = 0.0185).

نتیجه گیری: نتایج مطالعه حاضر نشان می دهد که EMQ از قابلیت اطمینان آزمون قابل قبولی برخوردار است. اکثر موارد (۸۰٪) از اصول MCQ پیروی می کنند. نتیجه می گیریم که EMQ می تواند به عنوان بخشی از ارزیابی تحصیلات تکمیلی برای آزمایش دانش مرتبه بالاتر و صلاحیت بالینی به طور مؤثر استفاده شود.

واژه های کلیدی: سؤالات تطبیق گسترده، تجزیه و تحلیل مورد، شاخص دشواری، صلاحیت بالینی

اینستھی سیا (بے ہوشی) کے رزیڈنٹس ای ایم کیو کے سوالات کی نفسیاتی جانچ پڑتال کی خصوصیات کا تجزیہ

بیک گراونڈ: کلینیکل تشخیص اطبا کے لئے نہایت ضروری صفت ہے ۔ پوسٹ گریجویٹ سطح پر کلینیکل تشخیص بنیادی حیثیت کی حامل ہوتی ہے۔ EMQ کے سوالات کلینکل سطح پر مرض کی تشخیص میں ممد و معاون ثابت ہوتے ہیں لیکن ان سے پوسٹ گریجویٹ سطح پر استفادہ نہیں کیا جاتا ہے ۔ کوویڈ ۱۹ کی عالم گیر وبا کی وجہ سے ڈاکٹر اور مریض کا براہ راست رابطہ کم ہوکر رہ گیا تھا اور میڈیکل اسٹوڈنٹس کی کلینیکل تعلیم میں ایک مسئلہ بن گیا تھا اسی وجہ سے کلینیکل تشخیص کے لئے نئی روشوں کو متعارف کرانے سے ڈاکٹروں کا کام آسان ہوجاتا ہے ۔ اسی وجہ سے اینستھی سیا کے رزیڈنٹ کے لئے EMQ کو متعارف کرایا گیا ہے تا کہ ان سوالات کی سائکو میٹریک خصوصیات سے آگہی حاصل کی جائے۔

روش: یہ تحقیقات ہندوستان کے شہر پانڈی چیری Pondicherry کے مہاتما گاندھی میڈیکل انسٹی ٹیوٹ میں انجام دی گئي ۔ ریزیڈنٹس کو ای ایم کیو کے چار ماڈل سوالنامے دئے گئے ۔ بیس ریزیڈنٹس کو چالیس تفصیلی سوالات کا ڈیٹا دیا گیا اور سائٹھ آپشنز دئے گئے تھے، سوالنامے کی علمی حیثیت آئٹم انالائسس اور کیوڈر رچرڈسن ۲۰ سے معین کی گئي ۔ DIF-I اور DIS-I کا بھی تجزیہ کیا گیا ۔

سفارش: اس تحقیق سے پتہ چلتا ہے کہ ای ایم کیو ایک قابل اطمینان ٹسٹ ہے ۔ البتہ اکثر موقعوں پر یعنی ۸۰ فیصد ایم سی کیو سے استفادہ کیا جاتا ہے۔ اس سے ہم نتیجہ حاصل کرتے ہیں کہ اس سے کلینیکل سطح پر تشخیص کی صلاحیتوں میں اضافہ ہوتا ہے اور اس سے کلینیکل سطح پر کافی فائدہ اٹھایا جاسکتا ہے ۔

کلیدی الفاظ: ای ایم کیو ، تجزیہ ، کلینیکل تشخیص

INTRODUCTION

Clinical competency is one of the most important attributes in postgraduate training as well as clinical reasoning is the core component of clinical competency. Pattern recognition, knowledge application, and intuitions are part of clinical reasoning (1). This makes it necessary to have an assessment system that is comprehensive, authentic and looks for the application of knowledge and not just factual recollection (2).

Assessment is the driving force for learning (3). It is one of the important influences on a student's learning experience and on the quality of learning. Every assessment method has its strength and weakness but its impact on the student's motivation and guidance to future learning is more valuable than these flaws (4). Miller's pyramid provides a conceptual framework for student assessment, including factual knowledge to problem-solving skills. Despite the availability of various tools for the formative assessment (FA) of postgraduates, most of the facilitators still practice traditional methods such as essay questions and case presentations. Although traditional assessment tools have good psychometric properties, their role in FA is minimal. Lack of effective feedback and scoring discrepancies are some of the pitfalls of traditional tools for FA (5). Newer methods of assessment such as Mini CEX, OSCE, Script concordance test, various forms of multiple-choice questions (MCQs) can be effectively used for FA of clinical reasoning.

Using MCQs could be the first step for the assessment of clinical competence (6). MCQs, in most situation, tests only factual knowledge and lower level of knowledge application, which is not sufficient for postgraduate assessment. To overcome the shortcomings of MCQ, extended matching questions (EMQ) were introduced. EMQs a variant of MCQ, is valid, feasible, and can be effectively employed for assessing problem-solving and clinical reasoning abilities in medical postgraduates (7,8). EMQs are less time consuming, easy to administer and can give immediate feedback. Despite these well-known advantages of EMQs, it is not commonly used in Postgraduate assessment. The Covid-19 pandemic hindered regular academic activities and also patient encounter for clinical training. This led us to introduce innovative teaching-learning and assessment strategy to sustain clinical reasoning and problem solving skills. Acknowledging the advantages of EMQs and its applicability during non-contact learning activities, we introduced EMQs among anaesthesia residents as part of the formative assessment and analysed its psychometric properties.

METHODS

After obtaining institutional ethics committee approval, the present item analysis of EMQs was conducted in the Department of Anaesthesiology, Mahatma Gandhi Medical College and Research Institute, Pondicherry. A series of EMQ tests were conducted in a "pen and paper" format as part of FA to our postgraduates from March 2020 to May 2020. The test was conducted among second and third-year residents only (n=20).

Construction and Pre-Validation of MCQs

All the EMQ items were based on the syllabus covered over the past six months. The items were prepared by two faculty trained appropriateness of the content, grammar, and construction. The EMQ test was divided into 4 modules; each module was based on a single theme which had a set of 10 questions and 15 options. The themes were a) Intra-operative critical incidents b) Post-operative complications c) Trauma and emergency d) Critical care management. A total of 40 clinical vignettes and 60 options were administered. Each correct answer was scored 1 mark and there was no negative marking for wrong answers. Cumulative score more than 50% was considered 'Pass'

Data processing:

For analysing the psychometric properties of the EMQs, the student's cumulative scores of all four modules were taken. Steps of item analysis:

1) The scores of all residents were arranged in descending order. 2) The upper third was considered a high achievers group **(HAG)** and the lower third as low achievers group **(LAG)** and the rest as Middle **(MAG)**. 3) Data of HAG and LAG were considered for analysis.

Item Analysis:

Item analysis involved 4 major parameter, i.e. test reliability, difficulty index (DIS-I), discrimination index (DIF-I) and distractor efficiency (DE). Reliability suggests whether an assessment tool is internally consistent and reproducible, it reflects the extent to which items within the test measures various aspects of the test. DIF-I shows whether the item was difficult or easy and how many students got the item correct. It ranged from 0 to 100%, where 0 indicated none of the students got the answer correct and 100% when all have got the item correct. DIS-I shows the ability of the item to differentiate a high achiever from a low achiever. It is expressed as a bi serial point correlation ranging from -1 to +1. Higher the index, better the item can differentiate achievers. Another important parameter was distractor functionality which is an independent indicator of the quality of an item. Distractors are said to be functional if it is selected by more than 5% of the students (9). Distractor efficiency is measured based on the number of non-functional distractors (NFD) in an item. It ranges from 0 to 100%. Items with 3, 2, 1 and 0 NFD had a distractor efficiency of 0, 33.3%, 66.6% and 100% respectively (10). Table 1 shows the interpretation of the item analysis parameters

Statistical analysis:

All data documentation and analysis was done using Microsoft® Excel (2013), IBM SPSS Statistics for Windows version 22, Armonk, NY: IBM Corp. Reliability co-efficient was calculated using the Kuder-Richardson 20 formula (KR-20). DIF-I and DIS-I were calculated and reported as mean and standard deviation. FD and NFD were expressed as percentages. The relationship between DIF-I and DIS-I was determined by Pearson correlation analysis and p-value of < 0.05 was considered statistically significant.

Table 1. Parameters of Item Analysis

Parameters	Formula	Interpretation of values
Reliability	Kuder Richardson 20	*Range = 0 to 1* 0= no reliability >0.7 = acceptable reliability 1 = high reliability
Difficulty Index (DIF-I)	(HAG+LAG) × 100 / N	*Range = 0 to 100%* < 30% = difficult 30% to 70% = acceptable >70% = easy
Discrimination Index (DIS-I	(HAG-LAG) × 2 / N	*Range = -1 to + 1* *(Bi-serial point correlation)* 0–0.19 = Poor discrimination 0.2 to 0.29 = Acceptable discrimination 0.3 to 0.39 = Good discrimination > 0.4 = Excellent discrimination
Distractor Efficiency (DE)	Percentages of non- functional distractors in an item(> 5% = functional distractor)	*Range = 0 to 100%* NFD 0= 100% NFD 1= 66.6% NFD 2= 33.3% NFD 3= 0%

***N** = Total number of students in both upper 1/3rd and lower 1/3rd groups, **HAG=** the number of students in upper 1/3rd group who answered correct, **LAG**= the number of students in lower 1/3rd group who answered correct. **NFD** = Non-functional distractor.

RESULTS

The present item analysis assessed 40 EMQs with 60 options divided into 4 modules, each module consisted of 10 EMQs with 15 options. The reliability coefficient of the test was 0.72. The mean test score was 24.5 ± 5.09, the highest score was 33 and the lowest was 11. Fifteen out of 20 residents had scored more than 50% marks (pass score). The mean test score according to the groups were: HAG 29.6 ± 2.5, MAG 23.8 ± 1.1, and LAG 19.6 ± 4.0, respectively.

Table 2 shows the overall analysis of the EMQ items. Table 3 shows distribution of items based on item difficulty and item discrimination. The mean difficulty index was 0.43 ± 0.17 with approximately 1/3rd of the items had moderate difficulty, 20% was difficult and only 3 items were easy. The mean item discrimination (0.28±0.24) with 5% negative discrimination, 20% of the items were having poor discrimination power (0-0.19), while 27.5% of the items exhibited excellent discrimination (>0.4). The remaining items were acceptable and good, out of which 25% of the items had an acceptable range (0.2 to 0.29) and 21% of the items showed good discrimination (0.3- 0.39). Figure 1 shows the distribution of DIF-I and DIS-I of the total EMQs, DIF-I with the lowest value of 0.05 and the highest was 0.74 whereas the lowest DIS-I value was -0.42 and the highest was 0.85.

Distractor analysis showed 6(10%) NFDs, 54(90%) were FD. Distribution of items based on the number of functional distractor showed, 13(32.5%) EMQs with one FD, 12(30%) items with two, and 6(15%) items with three FD.

The scattered diagram (Figure 2) represents the correlation between DIF-I and DIS-I of 40 items. The analysis showed a positive correlation between DIF-I and DIS-I (r = 0.215, p=0.018). The items on both poles of the difficulty index spectrum had poor or negative discrimination. Items with a moderate level of difficulty had exhibited good to excellent discrimination. Table 4 shows the relationship between DIF- I

Table 2. Descriptive analysis of EMQ items

Parameter	Mean (SD)
Total test score	24.3 ± 5.09
KR 20 (reliability coefficient)	0.72
DIF-I	0.43 ± 0.17
DIS-I	0.28 ± 0.24
Distractors	FD- 90% NFD- 10%

Table 3. Distribution of EMQs based on item difficulty and item discrimination

Parameter	Interpretation	EMQ items	
		N	%
Difficulty Index (DIF-I)			
< 0.3	Difficult	8	20%
0.3-0.7	Moderate	29	72.5%
0.7-1	Easy	3	7.5%
Discrimination Index (DIS-I)			
< 0	Negative	2	5%
0–0.19	Poor	8	20%
0.2 to 0.29	Acceptable	16	40%
0.3 to 0.39	Good	3	7.5%
> 0.4	Excellent	11	27.5%
Total		40	100%

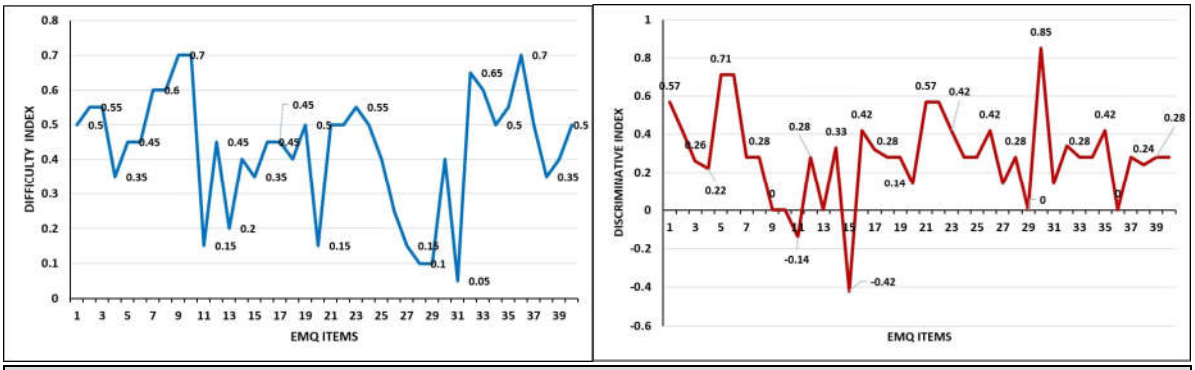

Figure 1. Composite of DIS-I and DIF-I of all EMQs

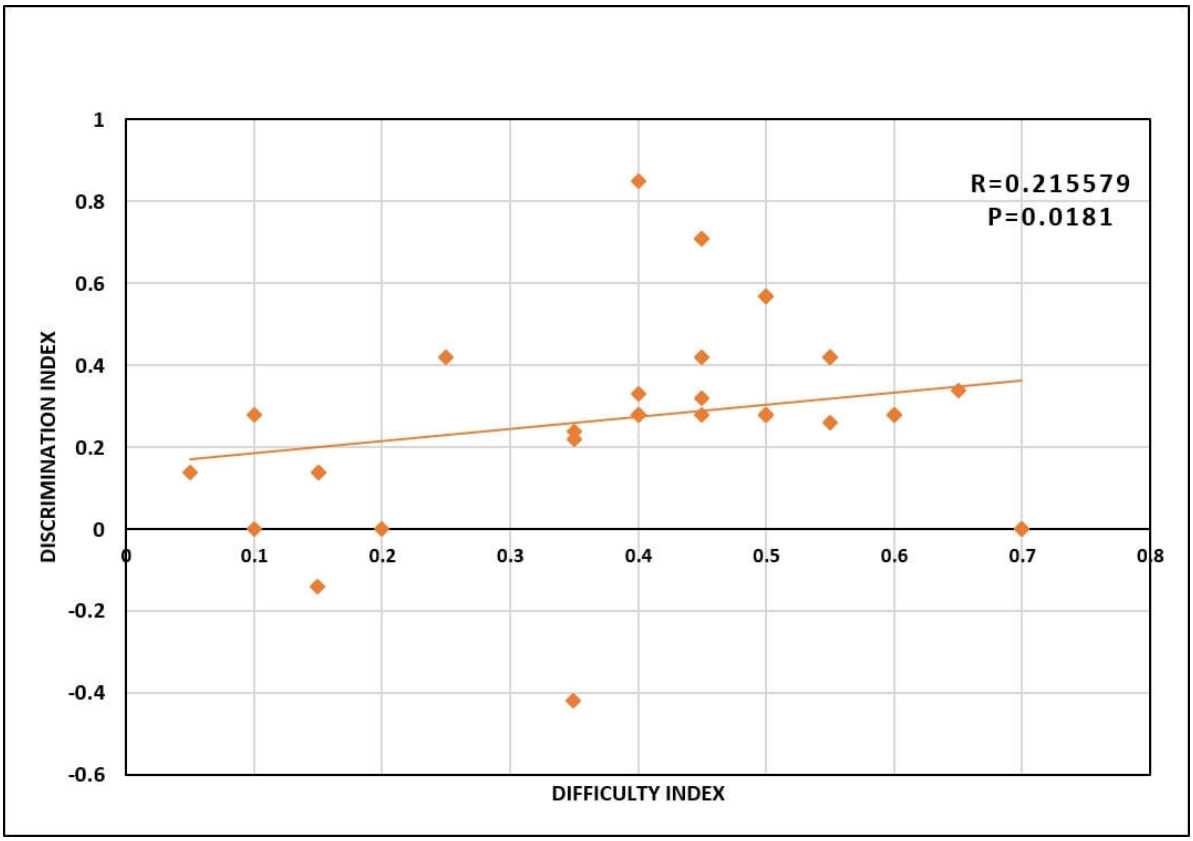

Figure 2. Correlation between DIS-I and DIS-I of EMQs

Table 4. Relationship between difficulty index and discrimination index of EMQs

DIS-I→ DIF-I ↓	Excellent >0.4	Good 0.3 – 0.39	Acceptable 0.2 - 0.29	Poor 0.0 – 0.19	Negative < 0
Acceptable (0.3-0.7)	10	3	15	0	1
Easy (0.7-1)	0	0	0	3	0
Difficult <0.3	1	1	0	5	1

and DIF-I for each item. Considering both DIFI and DIS-I together, 28 (10+3, 70%) EMQ items were acceptable. Another 2 (5%) EMQs, in spite of being difficult, were able to discriminate well between HAG and LAG. All easy items and 6 out of 8 difficult items were poor discriminators.

DISCUSSION

The present study focused on the application of extended matching questions for training and assessing clinical reasoning skills in anaesthesia postgraduates. This study showed an acceptable reliability coefficient of 0.72. Further analysis showed that 80% of the items had moderate difficulty and 70% items had good discrimination index. DIS-I and DIF-I correlated positively where, item discrimination increased with increase in item difficulty.

Clinical reasoning is one of the core features of clinical competency. There are various obstacles and challenges in integrating and assessing clinical reasoning in postgraduate training. FA should be able to capture the analytical capacity of the resident and identify areas for improvement. MCQs which are commonly used in medical education may not always assess higher-order thinking. Baig et al analysed MCQ in basic medical science exams and concluded that one-third of the MCQs tested recall of isolated facts while none of the MCQs assessed higher-order cognition (11). Recall of facts or eliminating incorrect answers is not a satisfactory way to master the subject. New assessment methods like script concordance test, key feature test, EMQs have shown good validity and reliability for assessing clinical reasoning. It can prevent answering by guess-work or pattern recognition (12). Item analysis is a quality control procedure for ensuring high-quality items. Similar to MCQs, EMQs have qualitative analysis for content, format, and writing procedures and quantitative analysis for psychometric properties.

The present item analysis assessed 40 EMQs with 60 options divided into 4 themes, each theme consisted of 10 EMQs with 15 options. This study showed a reliability co-efficient of 0.72 which is within the desirable range. The KR 20 Value of 0.7 was acceptable for a good assessment tool. Value close to 1 had higher homogeneity and consistency. Case et al showed that EMQs had higher reliability compared to an equal number of MCQ (0.55 vs. 0.42). A total of 52 EMQs are required to achieve a reliability of 0.72 and 105 EMQs for 0.85 (13).

The mean difficulty index of the EMQ items were 0.43±0.17, in the present study. The majority of the items of study had moderate difficulty and 20% of the items were very difficult. The DIF- I of this study is similar to other studies that evaluated EMQs and MCQs. Vuma S et al reported a mean DIF-I of ranged from 0.491 to 0.719 among three courses of EMQs conducted for third-year medical students (14). Keralia et al. reported mean DIF-I between 0.47–0.58 in MCQ items from 10 summative papers (15). Our analysis showed 7.5% EMQs were easy for both groups. Authors recommend the inclusion of items with all levels of DIF- I in a test but care should be taken to not compromise the quality of the paper. Edwardo Beckhoff set the median difficulty level at 0.5-0.6 with the following distribution: “easy items, 5%; items of medium-low difficulty, 20%; items of medium difficulty, 50%; medium-hard items, 20%; and difficult items, 5%” (16).

The mean DIS-I in the present study was 0.28 ± 0.24, where most of the items were good at discriminating high achievers from low achievers. Items with a high level of discrimination indices should be included in a test to enhance critical thinking. The authors recommend a DIS-I >0.2 to be included in the assessment. Similar to our results, studies have shown CPBR within 0.118 to 0.255 (14). The distribution of the DIS-I in the present study showed that 27% of the items had excellent and 40% had good discrimination. A total of 8 (20%) and 2 (5%) had poor and negative DIS-I in this study, respectively. Poor DIS-I can result in a low score due to a flaw in the items and needs to be removed from the bank. Too easy or too difficult items have poor discrimination which will decrease the reliability of the tool. These flawed items need to be reviewed for modification or discarded.

In the present study, item difficulty and item discrimination had a positive correlation ($r = 0.215$, $p=0.018$). The correlation was not linear, but more or less pyramidal/doom shaped. Several studies have also shown a dome-shaped correlation, items which are very easy or very difficult had poor discrimination capacity (17,18). Maximum DIS-I was seen in items with DIF-I between 0.3-0.7. Vuma et al also showed a positive correlation between DIF-I and DIS-I in EMQ tests (14). Contrary to our findings, Mitra et al, and Habib et al showed a significant negative correlation which indicates an inverse relationship between DIF-I and DIS-I: increase in difficulty index leads to a decrease in discrimination index (19). Our analysis combining the two indices (DIS-I and DIF-I) showed that 28 (70%) items could be called 'ideal' having a DIF-I between 0.3 - 0.7, as well as a DIS-I > 0.20.

Reducing the number of non-functional distractors and having more plausible options is one of the main aspects of a good quality item (20). Number of non-functional distractors in an item is inversely related to the discrimination power of an item and conversely higher distractor efficiency makes the item more difficult. Distractors of EMQs are not like MCQs, the pool of options applies to all the vignettes within the set. An answer to one clinical vignette can be a distractor for another vignette. Out of the 60 options, 54(90%) of the options were functional and only 6 (10%) options were selected by less than 5% of students. Previous studies have shown EMQs having 70 to 85% functional distractors and 14%-28% of NFD (14). The present also showed 1/3rd (13) of the items had one FD, another 1/3rd (13) had two FD. Several studies have debated the right number of options to create a good quality item. A large number of distractors decrease guesswork and increases reliability and validity. However, if a majority are non-functional in the option list it will only act as fillers and increase test time (21).

The overall analysis of the EMQ items showed that 31(77.5%) fulfilled the criteria for good quality items and can be retained for further use, while 2 items had negative discrimination which requires review. Items too easy or difficult with poor discrimination needed

to be discarded.

There are several limitations to this study. First, the total number of EMQs were small (n=40). But as a whole, the 40 EMQ items had good reliability (0.72) compared to studies which showed that 75 to 100 EMQ items are required to achieve a reliability > 0.75. Secondly, we did not elicit the perception of the student or faculty on the new format of FA. The study did not document the time taken to complete the test. Several studies have stated an increase in test time with an increase in the number of options. But in our study, the students completed the assessment well in time. Time taken for the EMQ test is much less compared to the conventional written assessment.

The present researchers concluded that EMQs have DIF-I and DIS-I within acceptable levels. Their continued use for assessing clinical reasoning in postgraduates as part of the FA is recommended. It is valid, reliable, and feasible and also has a good educational impact. Item analysis provides valuable information to improve reliability and validity. Faculty development and training in newer assessment tools like EMQs are required to prepare and bank quality items.

Ethical considerations

Ethical issues including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc. have been completely observed by the authors. The "Institutional Human Ethics Committee" MGMCRI, Sri Balaji Vidyapeeth approved this research and the ethics code is MGMCRI/IRC/04/2020/34/IHEC/186.

ACKNOWLEDGEMENT

The authors would like to thank Prof. Dr. Anantha Krishnan for his constant guidance, PGDHPE facilitators and batch mates for their support.

Financial Support: No external funds were received for this research.

Conflict of Interest: None to be declared.

REFERENCES

1. Thampy H, Willert E, Ramani S. Assessing Clinical Reasoning: Targeting the Higher Levels of the Pyramid. J Gen Intern Med. 34(8):1631-6.
2. Sood R, Singh T. Assessment in medical education: Evolving perspectives and contemporary trends. Natl Med J India 2012; 25:357-64.
3. Liu N, Carless D. Peer feedback: the learning element of peer assessment. Teaching in Higher Education. 2006; 11(3).279-90.
4. Epstein R. Assessment is medical education. NEJM. 2007; 356:387-96.
5. Miller GE. The assessment of clinical skills/competence/performance. Academic Medicine 1990, 65(9):563-7.
6. Tangianu F, Mazzone A, Berti F, Pinna G, Bortolotti I, Colombo F, et al. Are multiple-choice questions a good tool for the assessment of clinical competence in Internal Medicine? Italian Journal of Medicine 2018; 12:88-96.
7. Aisling SB. The new Extended Matching Question (EMQ) paper of the MFSRH Examination. Fam Plann Reprod Health Care. 2010; 36: 171-73.
8. Fenderson BA, Damjanov I, Robeson MR, Veloski JJ, Rubin E. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. Hum Pathol. 1997; 28(5):526-32.
9. Gronlund NE, Linn RL. Measurement and evaluation in teaching. 6th ed. New York: Macmillan publishing Co; 1990.
10. 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(1):17-20.
11. Eijsvogels T, Van Den Brand T, Hopman M. 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.
12. E J Wood. What are Extended Matching Sets Questions?. Bioscience Education. 2003; 1:1, 1-8
13. Kreiter CD, Ferguson K, Gruppen LD. Evaluating the usefulness of computerized adaptive testing for medical in-course assessment. Acad Med. 1999; 74:1125-8.
14. Vuma S, Sa B.A descriptive analysis of extended matching questions among third year medical students. Int J Res Med Sci 2017; 5:1913-20.
15. 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.
16. Backhoff E, Larrazolo N, Rosas M. The level of difficulty and discrimination power of the basic knowledge and skills examination (EXHCOBA). Rev Electro´n Investig Educ 2000; 2(1).
17. 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: 68-74.
18. Ramzan M, Imran SS, Bibi S, Khan KW, Maqsood I. Item Analysis of Multiple-Choice Questions at the Department of Community Medicine, Wah Medical College, Pakistan. Life and Science. 2020; 1(2): 60-63.
19. 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.
20. Namdeo SK, Sahoo S. Item analysis of multiple choice questions from an assessment of medical students in Bhubaneswar, India. Int J Res Med Sci 2016;4: 1716-19.
21. 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):S21-S24.