International Journal of Research in Medical Sciences
*Vuma S et al. Int J Res Med Sci. 2017 May;5(5):1913-1920*
www.msjonline.org
pISSN 2320-6071 | eISSN 2320-6012

DOI: http://dx.doi.org/10.18203/2320-6012.ijrms20171817

**Original Research Article**

# A descriptive analysis of extended matching questions among third year medical students

**Sehlule Vuma[1]*, Bidyadhar Sa[2]**

[1]Department of Para-clinical Sciences, Faculty of Medical Sciences, The University of the West Indies, St Augustine, Trinidad and Tobago
[2]Centre for Medical Sciences Education, Faculty of Medical Sciences, The University of the West Indies, St Augustine, Trinidad and Tobago

**Received:** 25 February 2017
**Accepted:** 25 March 2017

***Correspondence:**
Dr. Sehlule Vuma,
E-mail: Sehlule.Vuma@sta.uwi.edu



## ABSTRACT

**Background:** With changes in teaching methods in medicine, assessment tools have also evolved in order to be valid, reliable, practical, analyzable and not time-consuming. After questions of reliability and practicality when using free response short answer questions, we replaced them with extended matching questions (EMQs). Previous analysis of the same group of students, in the same time period, showed high reliability and discrimination with standard multiple choice questions (MCQ). Objective was to describe the efficiency of Extended matching questions (EMQ) in third-year medicine courses
**Methods:** Castler-Rock Integrity programme, item analyzed reports of EMQ results over a three-year period were analyzed. There were 25EMQ items in each course, each year, with 9 option answers.
**Results:** The Kuder Richardson-20 reliability mean ranged from 0.447 to 0.674. Spearman-Brown split half-reliability coefficient mean ranged from 0.443 to 0.685. Spearman-Brown prophecy reliability formula mean from 0.614 to 0 837. The Guttman split-half reliability coefficient mean ranged from 0.441 to 0.718. The difficulty mean ranged from 0.491 to 0.719. The Corrected point bi-serial coefficient ratio mean was 0.118 to 0.255. The number of items with all-functioning distractors ranged from 16% to 40%, and the total number of non-functioning distractors ranged from 14.5% to 28%.
**Conclusions:** EMQs showed reliability, though lower than with the MCQs previously analyzed. This may be due to the much smaller numbers hence increasing numbers of EMQs should be considered. There was a high number of functioning distractors. Poor distractors should be revised.

**Keywords:** Difficulty, Discrimination, Distractor, EMQ, Reliability

## INTRODUCTION

Questions requiring free-response, may have problems of reliability, validity, ambiguity, objectivity and practicability.[1] Answering and marking these open-ended questions is time-consuming compared to multiple choice questions (MCQs). There may also be discrepancies in marking. Sampling is also an issue whereas with MCQs more course content is examinable. Some writers however say, generally, standard MCQs assess factual knowledge rather than deeper understanding or use of information.[2] Others disagree: well written MCQs can assess higher level cognitive skills, although their creation requires more skill than basic recall type of questions.[3-5] Some schools have introduced extended matching questions (EMQs) to address some of these issues.

EMQs are variants of MCQs, with several advantages including:

- They have many of the advantages of MCQ tests objectivity, easy and quick to write and mark.[6-7]
- Depth and breadth of course content can be examined with less issues with sampling compared to SAQs.
- Item analysis can be performed to demonstrate reliability and validity.[8]
- Items do require application of knowledge and problem solving rather than simple recall of isolated pieces of information and basic facts.[6-7]
- They help to prevent students answering by elimination rather than actually knowing the answer.[6,9]
- They help in reducing the effect of "cueing".[9]
- The "theme" makes the examination content specific.[7]
- The structure itself facilitates item writing: the list of options flows easily and naturally.[7]
- Having homogeneous options reduces technical flaws.[7]
- The long option list allows inclusion of more possible options.[6]
- They can be changed to MCQs simply by decreasing the number of options[8] and the reverse is true.

The literature shows good levels of reliability with EMQs. Case and team compared MCQs with 5-answer options with EMQs with 9-23options on identical cases.[10] The generalizability coefficient of 18MCQs was 0.42, of 18EMQs was 0.55. Fenderson's team compared 20-option EMQs with 5-option MCQs: The internal consistency (Cronbach's alpha) of 240MCQs was 0.83 and of 220EMQs, 0.90.[11] On effective numbers of items, Kreiter et al showed that 52EMQs were required to obtain a reliability of 0.75 and 105 items for a reliability of 0.85.[12] Beullens et al showed reliability ranging from 0.73 to 0.86 for a series of EMQs.[7] Wass et al studied the construct validity of EMQs using.[13]

The correlation between EMQs and short answer questions (SAQs) was 0.60, 0.43 with true/false questions, -0.08 with an essay paper, 0.83 with an objective structured clinical examination (OSCE) and 0.48 with two long cases. They concluded that EMQs seemed to measure clinical problem solving, because the correlations with clinical tests (OSCE and long cases) and problem solving questions (SAQs) were high, moderate with a factual knowledge test (true/false questions) and low with a written presentation (essay). EMQs have indeed been used successfully by Licensing examinations like USMLE[11], Royal Australian and New Zealand College of Psychiatrists,[14] and MRCOG.[15] For MCQs, experts recommend wide ranges of item difficulties, ranging from difficult, to average and easy[5] with difficulty indices between 0.200-0.900.[16] For MCQ discriminators, most writers recommend discrimination coefficients of ≥0.20.[3] Some lower at 0.15 and others higher to 0.25.

This study was done among third-year medical students taking courses in para-clinical sciences, which integrate the sub-specialties of anatomical pathology, chemical pathology, haematology, immunology, microbiology, pharmacology and public health. the para-clinical sciences bridge the gap between the pre-clinical and clinical years. Teaching is a hybrid of problem based learning (PBL) and didactic lectures. Sub-specialties contribute equally in the combined examinations. Up until 2010, the final examination was a combination of MCQs and free response SAQ. Because SAQs were time consuming, and associated with marking discrepancies, they were replaced by EMQs. This paper analyzed the results of the newly introduced EMQs using item analysis. Item analysis strengthens question banks, and provides useful information for faculty to modify not only examinations but teaching methodologies.[17] Since EMQs engage students in higher level mental tasks, requiring application of knowledge and problem solving, they are in keeping with the PBL philosophy. MCQ Analysis of the same group of students, over the same period, by the same researchers showed high reliability and discrimination.[18]

### *Objective*

To describe the validity and reliability of newly introduced EMQs in 3, third-year medical school courses.

## METHODS

Approval was obtained from the Ethics Committee and the office of the Dean, Faculty of Medical Sciences. Item analysis of EMQ examinations of three courses, Course-I, Course-II and Course-III (C-I, C-II, C-III) in the academic years 2011-2012, 2012-2013, 2013-2014 was performed using the online Castler-Rock Integrity programme, each examination had 25EMQ items, each with 9 possible answer options: with instructions that "each option can be used once, more than once, or not at all" All examination papers were reviewed by the paraclinical sciences examination core-committee (for flaws, content and answer keys) and an external examiner before the students took the tests.

There was no negative marking for incorrect answers. Analysis included students' performance, reliability, using different indices including Kruder-Richardson-20 (KR-20), difficulty, discrimination using Corrected point bi-serial coefficient ratio (CPBR), distractors, and Pearson Correlations(r) between the different sub-specialties.[19] Further analysis of C-III (convenience) by sub-specialty was performed.

Poor item distracters (non-functioning) were those chosen by less than 5% of examinees. Three levels of difficulty were used: >0.75 (very difficult), 0.36-0.74 (moderate

difficulty), and ≤0.35 (low level difficulty). Four levels of item discrimination, using CPBR mean were used: ≥0.35 (high), 0.150-0.340 (good), 0.000-0.150 (poor), and <0 (negative) (no discrimination). Chi square ($\chi^2$) test of independence was calculated to assess the significance of the differences in different levels of difficulty and discrimination across all courses.

## RESULTS

Table 1 shows results of courses C-I, C-II, C-III. Students' mean scores were highest in C-III for all three years. There was moderate reliability by all indices. In C-II the indices were higher than C-I, and in C-III even higher. The item difficulty means were also higher in C-III.

**Table 1: Analysis of EMQs in years 2011-2012, 2012-2013, 2013-2014 in C-I, II and III.**

| z | 2011-2012 | | | 2012-2013 | | | 2013-2014 | | |
|---|---|---|---|---|---|---|---|---|---|
| **Course** | **C-1** | **C-II** | **C-III** | **C-I** | **C-II** | **C-III** | **C-I** | **C-II** | **C-III** |
| No. of students | 200 | 196 | 202 | 202 | 199 | 194 | 227 | 224 | 221 |
| No. of items | 25 | -25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 |
| Mean | 15.600 | 16.602 | 17.970 | 12.272 | 15.000 | 19.970 | 13.374 | 14.344 | 16.439 |
| Median | 16.600 | 17.000 | 18.000 | 12.000 | 15.000 | 18.000 | 13.000 | 14.544 | 15.000 |
| Mode | 14.000 | 18.000 | 21.000 | 14.272 | 14.000 | 21.000 | 13.000 | 15.000 | 16.000 |
| Standard deviation | 3.177 | 3.338 | 3.494 | 2.958 | 3.521 | 3.051 | 2.811 | 3.205 | 3.959 |
| Variance | 10.090 | 11.143 | 12.208 | 8.756 | 12.394 | 9.301 | 7.899 | 10.271 | 15.675 |
| Max score | 23 | 23 | 24 | 19 | 23 | 25 | 20 | 20 | 25 |
| Min score | 5 | 8 | 6 | 5 | 7 | 10 | 6 | 5 | 6 |
| Standard error of mean | 0.225 | 0238 | 0.246 | 0.208 | 0.250 | 0.219 | 0.187 | 0.214 | 0.266 |
| Standard error of measurement | 2.012 | 1.906 | 1.915 | 2.201 | 2.135 | 1.912 | 1.999 | 2.013 | 2.087 |
| KR-20- reliability | 0.599 | 0.674 | 0.699 | 0.447 | 0.632 | 0.607 | 0.494 | 0.605 | 0.722 |
| Spearman-Brown split half reliability coefficient | 0.601 | 0.667 | 0.685 | 0.443 | 0.634 | 0.612 | 0.480 | 0.608 | 0.719 |
| Spearman-Brown prophecy reliability formula | 0.751 | 0.800 | 0.813 | 0.614 | 0.776 | 0.760 | 0.648 | 0.756 | 0.837 |
| Guttman split-half reliability coefficient | 0.598 | 0.660 | 0.683 | 0.441 | 0.631 | 0.599 | 0.471 | 0.601 | 0.718 |
| Difficulty mean (range) | 0.624 (0.035-0.975) | 0.664 (0.031-0.980) | 0.719 (0.178-0.995) | 0.491 (0.025-0.955) | 0.600 (0.136-0.925) | 0.678 (0.031-1.000) | 0.535 (0.000-0.982 | 0.574 (0.067-0.951) | 0.658 (0.198 - 0.986) |
| CPBR mean (range) | 0.179 (-0.110-0.346) | 0.212 (-0.306-0.467) | 0.242 (0.006-0.436) | 0.118 (-0.054-0.328) | 0.203 (0.056-0.358) | 0.186 (0.048-0.363) | 0.134 (-0.055-0.310) | 0.179 (-0.242-0.393) | 0.255 (0.001-0.442) |

Table 2 shows number of items in the different levels of difficulty and discrimination: most falling in the moderately (range 32-72%) and very high difficulty (12-56%) over the three years. Most (range 40-68%) items had item discriminators above 0.15 range for the three courses. Those with negative CPBR were suppressed from the final results. Generally, the moderately difficult and the highly difficult items showed high discrimination. There were no statistically significant differences between the courses. Figure 1 shows examples of item discrimination by difficulty. The more difficult items showed better discrimination compared to easier items.

Table 3 shows distractor analysis. Total number of items with all functioning distractors over the three years ranged from 28% to 40%. Total number of non-functioning distractors ranged from 14.5% to 30%. Thus a high number of functioning distractors was achieved (70 to 85.5%). There was no statistically significant difference across the three years in different courses with regard to number of items with functioning and non-functioning distractors. However, with regard to total numbers of functioning and non-functioning distractors there exist significant differences for C-I and C-II ($P<0.01$). Tables 4 (2011-2012) and 5 (2013-2014) show results by the different sub-specialties for C-III. EMQs ranged between 2 and 6 items per sub-specialty. In 2011-2012 the item difficulties ranged from 0.178 to 0.995. The CPBR ranged from 0.006 to 0.436. In 2013-2014, item difficulties ranged from 0.198 to 0.986. The CPBR ranged from 0.001 to 0.442. In 2013-2014, the

correlations between the different sub-specialties ranged from small effect size to moderate range (Table 6).

**Table 2: Number of items in different levels of difficulty and discrimination.**

| Year | 2011-2012 | | | 2012-2013 | | | 2013-2014 | | |
|---|---|---|---|---|---|---|---|---|---|
| **Course** | **C-1** | **C-II** | **C-III** | **C-I** | **C-II** | **C-III** | **C-I** | **C-II** | **C-III** |
| **Difficulty** | | | | | | | | | |
| >0.75 | 12(48%) | 10 (40%) | 14 (56%) | 3 (12%) | 8 (32%) | *12 (48%) | 8 (32%) | 9 (35%) | 8 (32%) |
| 0-36-0.74 | 10 (40%) | 12 (48%) | 8 (32%) | 18 (72%) | 13 (52%) | 11 (44%) | 11 (44%) | 11 (44%) | 14 (56%) |
| ≤0.35 | 3 (12%) | 3 (12%) | 3 (12%) | 4 (16%) | 4 (16%) | 2 (8%) | **6 (24%) | 5 (20%) | 3 (12%) |
| $\chi^2$ correction df=4 | 1.121 P>0.05 | | | 5.749 P>0.05 | | | 0.686 P>0.05 | | |
| **Discrimination** | | | | | | | | | |
| ≥0.35 | 0 | 2 (8%) | 2 (8%) | 0 | 2 (8%) | 2 (8%) | 1 (4%) | 2 (8%) | 4 (16%) |
| 0.151-0.340 | 17(68%) | 16 (64%) | 17 (68%) | 11 (44%) | 18 (72%) | 11 (44%) | 10 (40%) | 16 (64%) | 17 (68%) |
| 0.000-0.150 | 6 (24%) | 6 (24%) | 6 (24%) | 11 (44%) | 5 (20%) | 12 (48%) | 12 (48%) | 4 (16%) | 3 (12%) |
| <0 (Negative) | 2 (8%) | 1 (4%) | 0 | 3 (12%) | 0 | 0 | 2 (8%) | 3 (12%) | 1 (4%) |
| $\chi^2$ correction df=6 | 1.443 P>0.05 | | | 7.342 P>0.05 | | | 8.729 P>0.05 | | |

* - One item had Difficulty of 1, **- One item had difficulty of 0.

**Table 3: Distractor analysis of C-I, C-II and C-III, in three years.**

| Course | C-I | | | C-II | | | C-III | | |
|---|---|---|---|---|---|---|---|---|---|
| **Year** | **2011-12** | **2012-13** | **2013-14** | **2011-12** | **2012-13** | **2013-14** | **2011-12** | **2012-13** | **2013-14** |
| No. of students | 200 | 202 | 227 | 196 | 199 | 224 | 202 | 194 | 221 |
| Total number of items | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 25 |
| No. of items with all functioning distracters | 9 (36%) | 10 (40%) | 7 (28%) | 12 (48%) | 9 (36%) | 4 (16%) | 9 (36%) | 10 (40%) | 10 (40.0%) |
| No. of items with non-functioning distracters | 16 (64%) | 15 (60%) | 18 (72%) | 13 (52%) | 16 (64%) | 21 (84%) | 16 (64%) | 15 (60%) | 15 (60.0%) |
| $\chi^2$ df=2 | 0.824 P>0.05 | | | 5.880 P>0.05 | | | 0.112 P>0.05 | | |
| Total number of distracters | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 |
| Total no. of non-functioning distracters (<5%) | 42 (21%) | 30 (15%) | 60 (30%) | 33 (16.5%) | 29 (14.5%) | 53 (26.5%) | 42 (21%) | 56 (28%) | 39 (19.5%) |
| Total no. of functioning distracters (>5%) | 158 (79%) | 170 (85%) | 140 (70%) | 167 (83.5%) | 171 (85.5%) | 147 (73.5%) | 158 (79%) | 156 (72%) | 161 (80.5%) |
| $\chi^2$ df=2 | 13.287 P<0.01 Significant | | | 10.671 P<0.01 Significant | | | 3.161 P>0.05 | | |

## DISCUSSION

### *Reliability*

Experts recommend high KR-20 which indicates reliable tests, suggesting an internally consistent instrument and showing test reproducibility and consistency.[20-22] A KR-20 closer to 1 is better at discriminating high performers from poorer performers. KR-20 of 0 does not show discrimination: meaning the item is easy.[20] Less than 0.3 is a poor discriminator.[20] Negative KR-20 shows unreliable tests.[21] Values of 0.7 are acceptable and for longer examinations e.g. with more than 50 items, KR-20 of 0.8 are desirable, Higher scores, >0.9, indicate that the examination is homogenous which is desirable. In the study KR-20 test means ranged from 0.447 to 0.722 demonstrating reliability. However it was lower in comparison with MCQs done by the same group of students in the same period which showed higher results (KR-20 test mean range 0.447 to 0.842).[18] This may be

due to the fact that the total number of EMQ items was low (25 compared to 75MCQs). This is even clearer with individual sub-specialty analysis. Kreiter et al showed that 52EMQ items were required to obtain a reliability of 0.75 and 105 items for a reliability of 0.85.[12] Indeed Wass et al suggested including more EMQs because they examine more of problem solving skills and were correlated more with other examinations that required problem solving skills like OSCE.[13] The reliability coefficients are better in the year 2013-2014. A possible explanation could be that staff are getting better at setting EMQs, or students are getting better at taking EMQ-format tests, or they are getting better at subject content.

**Table 4: EMQ: 2011-2012 Analysis of results by sub-specialty C-III.**

| | Anatomical pathology | Chemical pathology | Haematology | Immunology | Microbiology | Pharmacology | Total |
|---|---|---|---|---|---|---|---|
| No. of items | 5 | 4 | 3 | 5 | 4 | 6 | 25 |
| Max score | 5 | 4 | 3 | 5 | 4 | 6 | 24 |
| Min score | 0 | 1 | 0 | 0 | 1 | 0 | 6 |
| Mean | 4.317 | 3.035 | 2.287 | 3.015 | 3.317 | 3.832 | 17.970 |
| Std deviation | 0.919 | 1.029 | 0.783 | 1.340 | 0.857 | 1.151 | 0.2460 |
| Variance | 0.844 | 1.058 | 0.614 | 1.786 | 0.735 | 1.325 | 12.208 |
| SE mean | 0.065 | 0.072 | 0.055 | 0.094 | 0.060 | 0.081 | 0.246 |
| SE of measurement | 0.700 | 0.706 | 0.684 | 0.921 | 0.0694 | 0.0879 | 1.915 |
| KR-20 | 0.419 | 0.530 | 0.237 | 0.528 | 0.344 | 0.417 | 0.699 |
| Spearman-Brown split half reliability coefficient | 0.417 | 0.625 | 0.258 | 0.558 | 0.376 | 0.416 | 0.685 |
| Spearman-Brown prophecy reliability formula | 0.588 | 0.769 | 0.411 | 0.717 | 0.547 | 0.588 | 0.813 |
| Guttman split-half reliability coefficient | 0.391 | 0.579 | 0.248 | 0.532 | 0.363 | 0.406 | 0.683 |
| Skewness (total score) | -1.449 | -0.623 | -0.868 | -0.128 | -1.139 | -0.536 | -0.627 |
| Kurtosis (total score) | 2.370 | -0.894 | 0.126 | -1.017 | 0.542 | 0.381 | -0.056 |
| Difficulty range *(mean) | 0.767-0.970 | 0.574-0.995 | 0.629-0.837 | 0.302-0.842 | 0.738-0.936 | 0.178-0.936 | 0.178-0.995 (0.719) |
| CPBR range *(mean) | 0.074-0.434 | 0.003-0.516 | 0.037-0.180 | 0.070-0.477 | 0.126-0.206 | 0.079-0.327 | 0.006-0.436*(0.242) |

**Table 5: EMQ: 2013-2014 Analysis of results by sub-specialty: C-III.**

| | Anatomical pathology | Chemical pathology | Haematology | Immunology | Microbiology | Pharmacology | Total |
|---|---|---|---|---|---|---|---|
| No. of items | 5 | 5 | 2 | 5 | 6 | 2 | 25 |
| Max score | 5 | 5 | 2 | 5 | 6 | 2 | 25 |
| Min score | 0 | 0 | 0 | 0 | 1 | 0 | 6 |
| Mean | 2.321 | 3.452 | 1.240 | 0.083 | 1.004 | 0.344 | 16.439 |
| Std deviation | 1.240 | 1.399 | 0.759 | 1.325 | 1.115 | 0.674 | 3.959 |
| Variance | 1.537 | 1.958 | 0.576 | 1.755 | 1.243 | 0.454 | 15.675 |
| SE mean | 0.083 | 0.094 | 0.051 | 0.089 | 0.075 | 0.045 | 0.266 |
| SE of measurement | 1.004 | 0.091 | 0.0566 | 0.985 | 0.824 | 0.391 | 2.087 |
| KR-20 | 0.344 | 0.585 | 0.444 | 0.447 | 0.453 | 0.663 | 0.722 |
| Spearman-Brown split half reliability coefficient | 0.371 | 0.570 | 0.436 | 0.465 | 0.483 | 0.658 | 0.719 |
| Spearman-Brown prophecy reliability formula | 0.541 | 0.726 | 0.607 | 0.634 | 0.651 | 0.794 | 0.837 |
| Guttman split-half reliability coefficient | 0.360 | 0.552 | 0.436 | 0.445 | 0.461 | 0.657 | 0.718 |
| Skewness (total score) | 0.221 | -0.534 | -0.571 | -0.026 | -0.842 | -0.543 | -0.018 |
| Kurtosis (total score) | -0.536 | -0.790 | -1.052 | -0.786 | 0.067 | 0.944 | -0.677 |
| Difficulty range *(Mean) | 1.99-0.656 | 0.629-0.819 | 0.597-0.706 | 0.380-0.810 | 0.611-0.928 | 0.783-0.842 | 0.198-0.986 *(0.658) |
| CPBR range *(Mean) | -0.012-0.245 | 0.182-0.389 | 0.279-0.279 | 0.102-0.386 | -0.043-0.353 | 0.492-0.492 | 0.001-0.442 *(0.255) |

***Difficulty***

There was a range of difficulty as recommended by experts. The differences may also be due to the fact that some item constructors may be more advanced in this skill than others, or that some items examine easier course objectives. In C-III the difficulty mean showed more difficulty than in the C-I and C-II yet the students' mean scores were higher. This again may be because at this point students are at the end of the third year, are more comfortable with EMQs. C-III is in semester-2 and C-I is in the first half of semester-1 and C-II in the second half. Furthermore at this point, students have rotated through all clerkships. A lot more learning, from observation, occurs in clerkships where there is closer contact with staff in smaller groups and students are also exposed to practical and clinical application activities.[23-24]

**Table 6: EMQ: 2013-2014 C-III: Sub-specialty total score Pearson Correlation coefficients.**

| | Anatomical pathology | Chemical pathology | Haematology | Immunology | Microbiology | Pharmacology |
|---|---|---|---|---|---|---|
| Anatomical pathology | 1 | | | | | |
| Chemical pathology | 0.198 (p=0.003) | 1 | | | | |
| Haematology | 0.248 (p=1.905E-004) | 0.203 (p=0.002) | 1 | | | |
| Immunology | 0.308 (p=2.975 E-006) | 0.066 (p=0.329) | 0.26 (p=9.049 E-005) | 1 | | |
| Microbiology | 0.391 (p=1.672 E-009) | 0.289 (p=1.248 E-005) | 0.225 (p=7.725 E-004) | 0.238 (p=3.595 E-004) | 1 | |
| Pharmacology | 0.206 (p=0.002) | 0.076 (p=0.261) | 0.321 (p=1.079 E-006) | 0.297 (p=7.083E-006) | 0.24 (p=3.136 E-004) | 1 |

***Discrimination***

To promote and enhance critical thinking, items need to have high levels of discrimination power.[25] For MCQs, writers recommend discrimination coefficients of ≥0.20.[3] Some may go as low as 0.15 and others higher to 0.25. In the study most items were good at discriminating high performers from the poorer performers. CPBR ranged from -0.306 to 0.467 mostly >0.15. Items with negative CPBR were discussed with relevant staff and were suppressed from the final students' results. Ware J et al created arbitrary levels of discrimination power where >0.4 was excellent, 0.30-0.39 was good, and 0.15-0.29 was moderate and below 0.15 was considered to have poor discrimination power.[26] They showed over 4 years the excellent category ranged from 0.8% to 21% and the very good category to range between 10 to 19%.

***Distractors***

Discriminating powers increase with increasing numbers of functioning distractors.[3] There were high numbers of functioning distractors. The number of items with all-functioning distractors ranged from16 to 40% and in comparison, the same group of students over the same period, in the MCQ analysis showed total number of MCQ items with all functioning distractors ranged from 34.9 to 65.3%.[18] With more answer options, (9 compared to 4), it may be more difficult to create plausible alternatives that draw students to respond to them. The total number of non-functioning distractors ranged from 14.5 to 28%: compared to MCQs which ranged from 14.2% to 36.8% in MCQs: which was comparable.[18]

Tarrant M et al showed only 13.8% items had functioning distracters in 4 or 5 option-MCQs stating that some teachers have problems constructing good distracters.[27] They emphasized that the important thing was the quality of distracters and not quantity: even suggesting decreasing to three options. Some researchers however argue that 3 options increase chances of students just guessing. More distractors decrease the probability of this guess work, and increases reliability and validity.[28,29] At the same time, however, increasing the options increases the test-time.[29] Furthermore, good quality, well-constructed distractors, does reduce cueing.[30] Hence the importance of EMQs with more distracters. However distractors can't just be "fillers" they need to be good and plausible. Non-functional distracters just increase test time unnecessarily.[31] Previously research showed more, plausible, options made the examination harder, more discriminating, and 8 options were showed better precision and better testing times.[31]

***Correlations***

Correlations were positive among all sub-specialties suggesting that the sub-specialties were well aligned, however the alignment was stronger with the previously analyzed MCQs which ranged from 0.208 to 0.476.[18] This again may be a question of smaller numbers of EMQs. Earlier, the same researchers showed correlations

between MCQs, free response progressive disclosure questions (PDQ), and the clinical/practical based Objective structured practical examinations (OSPE), to range between 0.208-0.354 for the haematology component of the integrated examinations.[24]

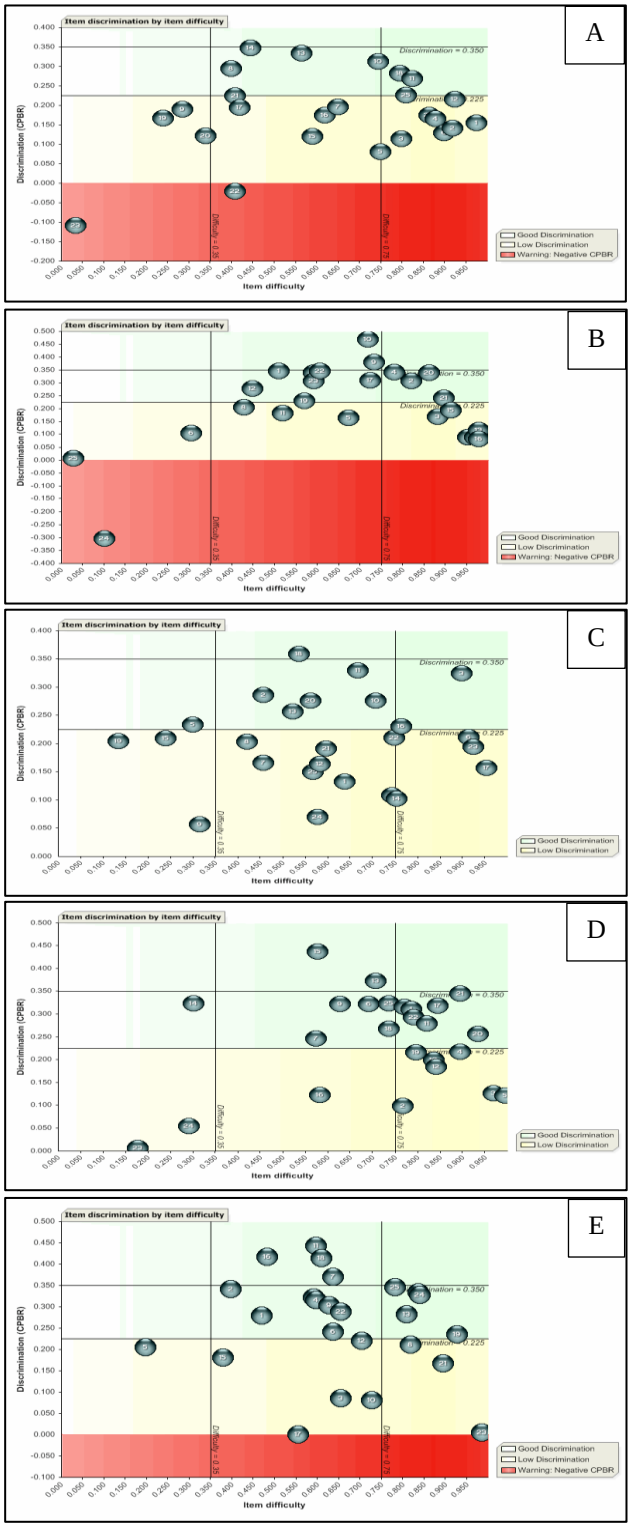


**Figure 1: Examples of Item difficulty by item discrimination: C-I, CII and C-III. A) CI: 2011-2012; B) CII: 2011-2012; C) CII: 2012-2013; D) CIII: 2011-2012; E) CIII: 2013-2014.**

***Limitations***

- There was a small number of test items especially when broken down by sub-specialty.
- The subspecialty analysis was done only in C-III for convenience, but all because this course showed the best results.
- The study did not document the views of students nor staff on EMQs: however there being no need for marking SAQ scripts left a lot of time to plan for the next academic year. The study did not document the time taken to answer EMQs since reports say that increasing item options increases the test-time.[31] However all students finished the examinations well within time (This was a paper-based examination).

## CONCLUSION

EMQs had acceptable levels of difficulty, discrimination and distracters. Their continued use is recommended, however to improve reliability the total number of items should be increased, perhaps to 75 like MCQs. More course content can be examined too. Well-constructed EMQs are able to assess higher order knowledge and skills like application of basic knowledge and problem solving skills. Data from item analysis are very valuable and training sessions for item writing are recommended to improve quality.[26] The items with many poor distractors should be revised, or even be converted to MCQs instead.

## ACKNOWLEDGEMENTS

Authors would like to thank Former Dean, Prof Samuel Ramsewak, Mrs. Louise Green former staff of the Assessment unit and Centre for Medical Sciences Education (CMSE), staff of the department of Para-clinical Sciences, Faculty of Medical Sciences, University of The West Indies, St Augustine, Trinidad and Tobago.

*Funding: No funding sources*
*Conflict of interest: None declared*
*Ethical approval: The study was approved by the Institutional Ethics Committee*

## REFERENCES

1. Case SM, Swanson DB. Extended matching items: a practical alternative to free-response questions. Teaching and Learning in Medicine. 1993;5:107-15.
2. Azer SA. Assessment in problem-based learning. Biochem Mol Biol Educ. 2003;31(6):428-34.
3. DiBattista D, Kurzawa L. Examination of the quality of multiple-choice items on classroom tests. Canadian J Scholarship Teaching and Learning. 2011;2(2):4.
4. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions?

Research Paper. BioMed Central Medical Education 2007;7:49
5. Campbel DE. How to write good multiple-choice questions. J Paediatr Child Health. 2011;47:322-5.
6. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. Third edition (Revised) National Board of Medical Examiners, 3750 Market Street, Philadelphia: PA; 2002:19104.
7. Beullens J, Damme BV, Jaspaert JJ, Janseen PJ. Are extended-matching multiple-choice items appropriate for a final test in medical education? Medical Teacher. 2002;24(4):390-5.
8. Wood EJ. What are extended matching sets questions? BEE-J. 2003;1(1):1-9.
9. Epstein RM. Assessment in medical education. N Engl J Med. 2007;356:387-6.
10. Case SM, Swanson DB, Ripkey DR. Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. Acad Med. 1994;69(10):S1-3.
11. 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;28:526-32.
12. Kreiter CD, Ferguson K, Gruppen LD. Evaluating the usefulness of computerized adaptive testing for medical in-course assessment. Acad Med. 1999;74:1125-8.
13. Wass V, Mcgibbon D, van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? Med Educ. 2001;35:326-30.
14. Samuels A. Extended matching questions and the Royal Australian and New Zealand College of Psychiatrists written examination: an overview. Australas Psychiatr. 2006;14(1):63-6.
15. Duthie S, Fiander A, Hodges A. EMQs: a new component of the MRCOG Part 1 examination. Obstetric Gynaecol. 2007;9:189-94.
16. Medical Council of Canada. February 2010. Guidelines for the Development of Multiple-Choice Questions. (http:// mcc.ca/ wp-content/ uploads/ Multiple-choice-question-guidelines.pdf).
17. 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;4:1716-9.
18. 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.
19. Pearson's r Correlation. (Accessed September 26, 2015). Available at: http:// faculty.quinnipiac.edu/ libarts/posci/Statistics.html
20. KR-(20). (Accessed September 26, 2015). Available at: http://eacvisualdata.com/eacs/kr20.aspx.
21. Thompson NA. (Accessed August 21, 2015). KR-20. Available at: http:// knowledge.sagepub.com/ view/researchdesign/n205.xml
22. Kuder and Richardson Formula 20. (Accessed September 26, 2015). Available at: http://www.real-statistics.com/reliability/kuder-richardson-formula-20/
23. Vuma S, Sa B, Ramsewak S. Descriptive analysis of pre-testing outcome in haematology as an indicator of performance in final examinations among third year medical students. Caribbean Teaching Scholar. 2015;5:1:25-35.
24. Vuma S, Sa B, Ramsewak S. A retrospective co-relational analysis of students' performance in different modalities of assessment in Haematology and the final integrated multi-specialty examinations among third year MBBS students. Caribbean Teaching Scholar. 2015;5:1:37-46.
25. Morrison S, Walsh K. Writing multiple-choice test items that promote and measure critical thinking. J Nurs Educ. 2001;401.
26. Ware J, Vik T. Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. Med Teach. 2009;31(3):238-43.
27. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distracters in multiple-choice questions: a descriptive analysis. BMC Medical Education. 2009;9:40.
28. Considine J, Botti M. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. Collegian. 2015;12:1.
29. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-Choice Item-Writing Rules. Applied Measurement in Education. 1989;2(1):51-78.
30. Hift RJ. Should essays and other open-ended- type questions retain a place in written summative assessment in clinical medicine? BMC Medical Education. 2014;14:249.
31. 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. J Assoc Am Coll. 2008;83(10):S21-4.

**Cite this article as:** 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.