Journal of Taibah University Medical Sciences (2017) 12(1), 14–26

Taibah University

Journal of Taibah University Medical Sciences

www.sciencedirect.com

Original Article

A comparison of clinical-scenario (case cluster) versus stand-alone multiple choice questions in a problem-based learning environment in undergraduate medicine

Sehlule Vuma, FCPath[a,*] and Bidyadhar Sa, PhD[b]

[a] *Department of Para-clinical Sciences, Eric Williams Medical Sciences Complex, Faculty of Medical Sciences, The University of the West Indies, St Augustine, Trinidad and Tobago*
[b] *Centre for Medical Sciences Education, Eric Williams Medical Sciences Complex, Faculty of Medical Sciences, The University of the West Indies, St Augustine, Trinidad and Tobago*

Received 30 July 2016; revised 24 August 2016; accepted 28 August 2016; Available online 11 November 2016

الملخص

أهداف البحث: المقارنة بين عناصر الأسئلة متعددة الاختيار ”القائمة بذاتها“ وبين عناصر أسئلة ”السيناريو السريري“ المدمجة متعددة الاختيار (الحالة المتراكبة) في بيئة التعلم القائم على حل المشاكل.

طرق البحث: تم إجراء تحليل وصفي استرجاعي على امتحانات متعددة الاختيار في مقرر يدمج تخصصات علم الأمراض التشريحي، وعلم الأمراض الكيميائي، وعلم أمراض الدم، وعلم المناعة، وعلم الأحياء الدقيقة وعلم الصيدلة. تم تحليل عناصر الأسئلة متعددة الاختيار من ناحية الموثوقية (كودر-ريتشاردسن-٢٠) ومستوى الصعوبة، ومؤشر التميّز، وصارفات الانتباه عن العنصر وأداء الطلاب. استخرجت التحاليل الإحصائية للنتائج من برنامج تحليل سلامة العنصر المتوفر على الانترنت. تم بعد ذلك مقارنة نتائج الأسئلة متعددة الاختيار ”القائمة بذاتها“ مع نتائج أسئلة ”السيناريو السريري“ المدمجة متعددة الاختيار (الحالة المتراكبة).

النتائج: كانت نتائج كودر- ريتشاردسن-٢٠ بالنسبة لأسئلة ”السيناريو السريري“ المدمجة متعددة الاختيار أعلى من نتائج الأسئلة متعددة الاختيار ”القائمة بذاتها“ على الدوام. وكانت قيم كودر- ريتشاردسن-٢٠ ومستوى الصعوبة أعلى بالنسبة لأسئلة ”السيناريو السريري“ المدمجة متعددة الاختيار. بالنسبة لمستوى الصعوبة ومؤشر التميّز، لم يكن هناك فرق ذو قيمة إحصائية بين أسئلة ”السيناريو السريري“ المدمجة متعددة الاختيار والأسئلة متعددة الاختيار ”القائمة بذاتها“. وكان هناك مجموعة من مستويات الصعوبة على تصنيف بلوم. كما أن متوسط درجات الطلاب كان أعلى في امتحانات السيناريو السريري المدمجة متعددة الاختيار. وكان إعداد امتحانات ”السيناريو السريري“ المدمجة متعددة الاختيار أكثر تحديا.

الاستنتاجات: تضاهي امتحانات ”السيناريو السريري“ المدمجة متعددة الاختيار امتحانات الأسئلة متعددة الاختيار ”القائمة بذاتها“ وتوفر فرصا للدمج التكاملي بين التخصصات الفرعية والتقييم المتماشي مع طريقة التعلم المبنية على حل المشاكل. حيث تقوم بتقييم المهارات المعرفية للطلاب مع كونها موثوقة وعملية. وتشجع المستويات المختلفة من صعوبة العناصر فيها التفكير النقدي ومتعدد المنطق. كما كانت درجات الطلاب أعلى في امتحانات السيناريو السريري المدمجة متعددة الاختيار، الذي قد يشير إلى فهم أفضل للمادة، أو وضوح أكثر في السؤال. وينبغي على السيناريوهات أن تتتابع بصورة منطقية. كما أن زيادة عدد السيناريوهات يضمن فحصا أشملا لمحتوى المقرر.

الكلمات المفتاحية: السيناريو السريري؛ الصعوبة؛ التمييز؛ الدمج؛ التعلم المبني على حل المشكلة

Abstract

Objectives: To compare stand-alone multiple choice questions (MCQs) and integrated clinical-scenario (case cluster) multiple choice questions (CS-MCQs) in a problem-based learning (PBL) environment.

Methods: A retrospective descriptive analysis of MCQ examinations was conducted in a course that integrates the subspecialties of anatomical pathology, chemical pathology, hematology, immunology, microbiology and pharmacology. The MCQ items were analyzed for their reliability (Kuder–Richardson-20, KR-20), level of difficulty (Pi), discrimination index (Di), item distractors and student performances. The statistical analysis of the

* Corresponding address: Department of Para-clinical Sciences, Eric Williams Medical Sciences Complex, Faculty of Medical Sciences, The University of the West Indies, St Augustine, Trinidad and Tobago.
E-mail: Sehlule.Vuma@sta.uwi.edu (S. Vuma)

Peer review under responsibility of Taibah University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jtumed.2016.08.014

results was extracted from the integrity online item-analysis programme. The results of the standard stand-alone and CS multiple choice questions were compared.

Results: KR-20 for the CS-MCQs and stand-alone MCQs was consistently high. KR-20 and Pi were higher for the CS-MCQs. There was no significant difference between the CS-MCQs and stand-alone MCQs in Pi and Di. A range of difficulty levels was found based on Bloom's taxonomy. The mean scores for the class were higher for the CS-MCQ examination. The compilation of the CS-MCQ examination was more challenging.

Conclusions: CS-MCQs compare favorably to stand-alone MCQs and provide opportunities for the integration of sub-specialties and assessment in keeping with PBL. They assess students' cognitive skills and are reliable and practical. Different levels of item difficulty promote multi-logical and critical thinking. Students' scores were higher for the CS-MCQ examination, which may suggest better understanding of the material and/or better question clarity. The scenarios have to flow logically. Increasing the number of scenarios ensures the examination of more course content.

Keywords: Clinical scenario; Difficulty; Discrimination; Integration; PBL

Introduction

Problem-based learning (PBL) is one of the most accepted modes of curriculum delivery in medical schools.[1] It discourages students from simply obtaining basic factual knowledge[2] and encourages and emphasizes the integration of basic knowledge and clinical skills. One challenge for teachers is to design assessment strategies that are in line with the PBL philosophy.[1] Assessments should match the competencies that the students are to learn and the teaching format used.[1]

Currently, multiple choice question (MCQ) examinations are a widely accepted assessment modality. Convincing evidence by researchers shows that MCQs not only satisfy all psychometric characteristics (reliability, validity, objectivity, fairness and practicality) of testing but also assess higher-order thinking with precision. Practicality in terms of both human and material resources in planning and implementing a test is very important.[7] Some writers support the use of MCQs, whereas others[2] are of the view that for the most part, standard MCQs assess only factual knowledge or the use of information rather than deeper understanding of content or cognitive skills; thus, they are not always useful for PBL assessment.

Other authors state that well-written MCQs do assess higher-level cognitive skills, although creating these items requires more skill than the basic recall type of questions.[3,4] PBL content assessment using MCQs in combination with computer-based objective tests (COMBOT) was shown to be significantly reliable and well aligned with the major learning outcomes of PBL cases.[5] Essays or short answer questions (SAQs), while they may address deeper thinking and higher cognitive level skills, are time consuming and are associated with grading discrepancies and variations.[3] They are more difficult to grade.[8] The modified essay question (MEQ) examination, also known as progressive disclosure questions (PDQs), was introduced as a compromise between the essay/SAQ and MCQ.[3] However, some authors have shown that, while the intent was indeed to ask questions requiring higher-order cognitive skills, the PDQ examination questions actually required predominantly lower-order cognitive skills.[6,9] Some schools have introduced extended matching questions (EMQs) and others clinical scenario MCQs (CS-MCQs) (also known as "case clusters").[2,10–12]

CS-MCQs assess students in a similar way as MEQ/PDQs. In MEQ/PDQs, a clinical case is given and questions are asked based on the case. Each question may reveal further information progressively as required.[3] They test analytical skills, problem solving skills, cognition and the integration of knowledge. They encourage students to think not just about basic knowledge or individual systems but about the whole patient,[3] which better reflects the learning process[11] and also better prepares students to assess their patients when they become doctors in the future.[11] Further, compared to MEQ/PDQs, they have all the advantages of MCQs. They are easy and less time consuming for staff to grade and less time consuming for students to write. They examine more course-content in a short time, and have fewer problems associated with sampling as observed in MEQs/PDQs.[9] Indeed some researchers[6] have shown more item flows with MEQs than with MCQs.

When comparing MCQs preceded by clinical scenarios and exact items (based on the same exact topics), it was shown that while the time required to answer CS-MCQs increased by 20%, students perceived that in the integrated course, the clinical scenarios improved question clarity and increased relevance to the curriculum.[11] CS-MCQ tested the students' ability to synthesize information as well their clinical reasoning.[10] Indeed, medical education experts Case and Swanson in 2002 agreed that case-clusters are particularly important for PBL courses because they test the application of knowledge.[12] However, it is important in this format to be careful and avoid "cueing and hinging"[12]: no "hinging" unless the topic is so important that it is an "all or nothing".[12]

Quality control exercises are important for ensuring high-quality MCQs.[13] MCQ items can be analyzed qualitatively (for content validity, form, and effective writing procedures) and quantitatively (for statistical properties, which include a measurement of item difficulty (Pi), the item discrimination index (Di) and item distractors). MCQ items should be modified to have Pi and Di within acceptable ranges.[14] Effective items discriminate between high and low scorers throughout the test. Ideal items have the most high scorers passing and low scorers failing.[15–17]

Objective: To compare stand-alone MCQ and CS-MCQ items over three years in a third year undergraduate medical course, in the department of Para-clinical Sciences, in a PBL environment.

Materials and Methods

Setting

Para-clinical Sciences integrate the sub-specialties of anatomical pathology, chemical pathology, hematology, immunology, microbiology, pharmacology, and public health. Teaching is a hybrid of didactic lectures and PBL and is systems based. PBL is more of a "Guided Discovery Approach", as opposed to an "Open Discovery approach". Students rotate through all sub-specialties in clerkships throughout semesters 1 and 2. The courses Applied Para-clinical Sciences-I (APS-I) and Applied Para-clinical Sciences-II (APS-II) are in Semester 1, and Applied Para-clinical Sciences-III (APS-III) is in Semester 2. For APS-I, II, and III, PDQs and PBL-tutor assessments are used for in-course/formative assessments, and MCQs/EMQs are used for end of course/summative assessments. The introduction of PDQs in 2009 provided an opportunity to have an examination that integrated all the sub-specialties. An analysis of the PDQ examination showed more than 50% were basic level questions,[18] similar to what was shown elsewhere.[6,9] Also similarly noted was the issue of under and over representation of some sub-specialties due to sampling. Not all sub-specialties will have relevant objectives in every given clinical scenario. APS-I and II use the standard stand-alone MCQs. CS-MCQs were introduced in 2009 to only APS-III, integrating the sub-specialties in a similar way as PDQs because PBL encourages integration.

Ethical approval was obtained from the Ethics Committee and the Office of the Dean, Faculty of Medical Sciences. The study was a retrospective descriptive analysis conducted from February to September, 2015. Students' performance on the MCQ examinations in APS-III for the academic years 2011—2012, 2012—2013, and 2013—2014 was analyzed. The MCQ items in APS-III for the same academic years were analyzed for reliability, Pi, Di, and item distractors. The statistical analysis of the results was extracted from the integrity online item analysis programme (http://integrity.castlerockresearch.com/About.aspx). The results of the standard stand-alone MCQs and integrated CS-MCQs were compared.

1: Three levels of Pi were used: >0.75 (very difficult), 0.36—0.74 (moderate difficulty), and ≤0.35 (low level difficulty).

2: Five levels of Di, using the corrected point-biserial ratio (CPBR) mean, were used: ≥0.35, 0.226—0.340, 0.160—0.225, 0.000—0.150, and <0 (Negative). A CPBR of ≥0.35 was considered high and a negative CPBR was considered very poor. These items were removed from the final student results.

3: Items were assigned a cognitive level (by the authors/researchers), based on the level of Bloom's taxonomy of the objectives that the questions required of the students.[19] (Bloom's taxonomy was modified and assigned based on whether the students were being asked for the basic recall of simple facts, e.g., Level I, where the instructional verb of the objective was "list/name"; Level II, the recall of more difficult facts and comprehension where the instructional verb was "explain/describe" e.g., concepts, mechanisms, and pathogenesis; Level III, the comprehension and application of basic facts in a clinical scenario; and Level IV, problem solving and interpreting e.g. sets of results or clinical presentation and suggesting further investigations and expected results, management, complications etc.) The Chi square (χ^2) test of equality for the percentage of questions in each level was used to assess the significance of the differences in the distribution across the four levels of I, II, III and IV.

4: Poor item distractors (non-functioning) were those chosen by less than 5% of the examinees.

Because of the small numbers of stand-alone MCQs in APS-III, the results of APS-I and APS-II (which are stand alone with no case clusters and no integration between sub-specialties) were also analyzed for comparison.

Results

The majority of the integrated CS-MCQs involved two to six sub-specialties. A few involved just one sub-specialty. The total number of items per scenario ranged from 2 to 8. In 2011—2012 and 2012—2013, two case clusters were hinged each year (range of items per cluster was 2—3). In 2013—2014, there were five such case clusters, (range of items per cluster also 2 to 3). Items had one correct option and three distractors, and no negative marking was used.

Figures 1—3 show the results of Di by Pi for APS-III. The moderately and highly difficult items show higher discrimination than the easier items. Table 1 shows the results of the integrated CS-MCQs versus the stand-alone MCQs in APS-III (Pi, CPBR, (Di), items by Bloom's taxonomy levels), and KR-20. Table 2 shows the Chi square (χ^2) statistics. Statistically, except for 2011—2012 and 2013—2014 for Pi and 2013—2014 for Di, there is no significant difference between the CS-MCQs and stand-alone MCQs. Table 3 shows the analysis of item distractors. For all three years, there were no statistically significant differences between the CS-MCQs and stand-alone MCQs with regards to the number of items with all-functioning distractors and non-functioning distractors. Table 4 shows the integrity analysis of the three years for all three courses APS-I, II and III: (students' performance, Di, Pi, reliability, test (Kuder—Richardson-20) (KR-20)). In Table 5, in the three courses, between the years, there are no significant differences in the KR-20 reliability coefficient. Table 6 shows the item distractors in the CS-MCQs against the stand-alone MCQs in APS-I, APS-II and APS-III over the three years. The major difference was in APS-II in 2012—2013 (non-functioning distractors made up only 14.2%). There were no statistically significant differences across the years in APS-I and APS-III with regards to the number of items with all-functioning distractors and non-functioning distractors. The correlations (Table 7) between the different sub-specialties were mostly of the medium effect size range. An example of the analysis of two of the CS-MCQs (with the

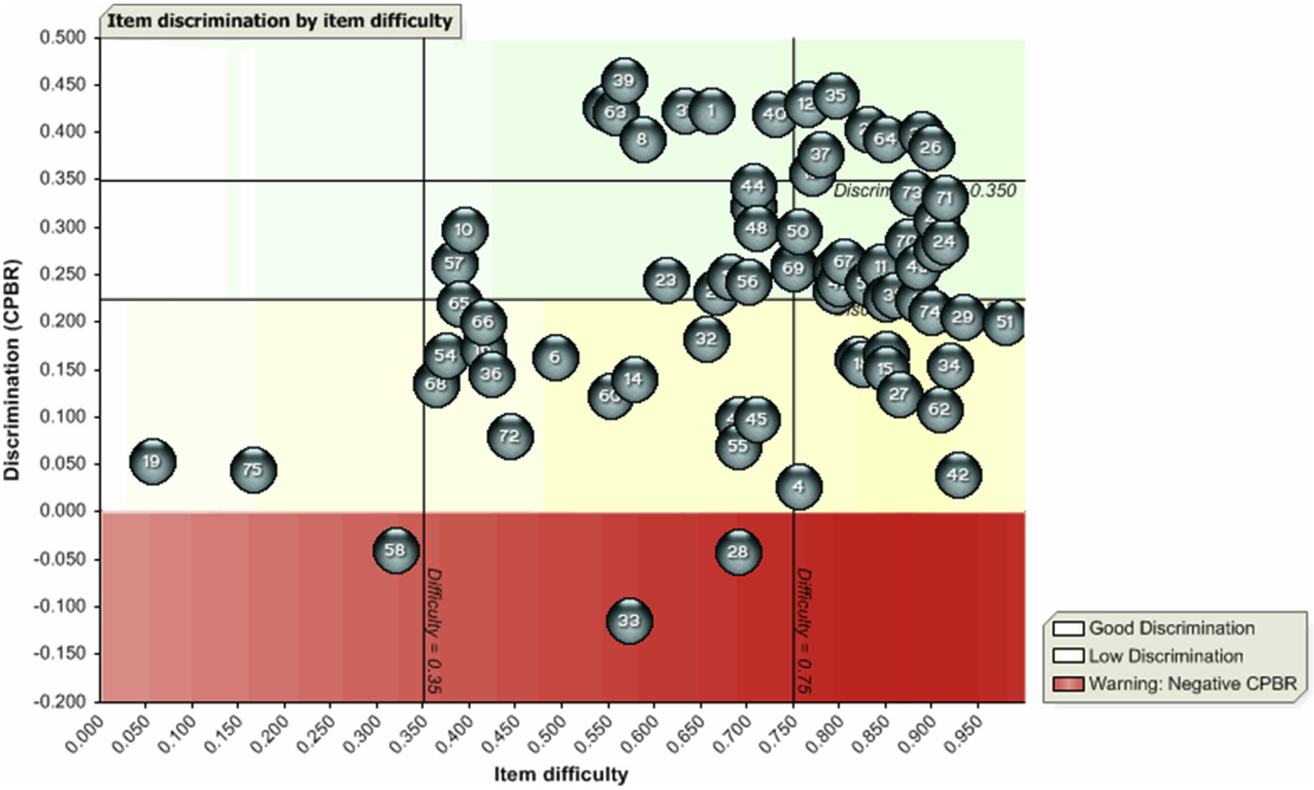

Figure 1: APS-III: Item Discrimination/Difficulty: Year 2011−2012 (Case clusters-item1-67, stand-alone: item 68−75).

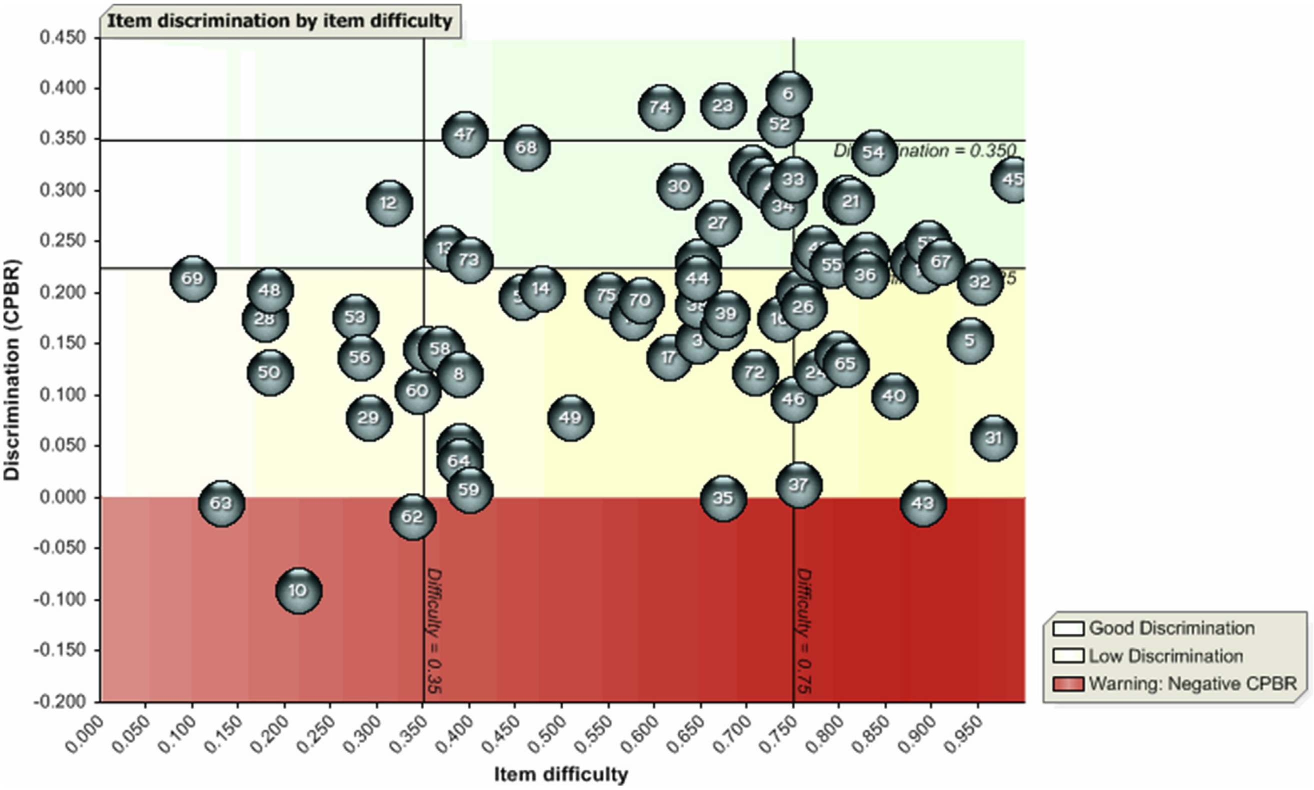

Figure 2: APS-III: Item discrimination/difficulty: Year 2012−2013 (Case clusters-Item 1−63, stand alone: Item 64−75).

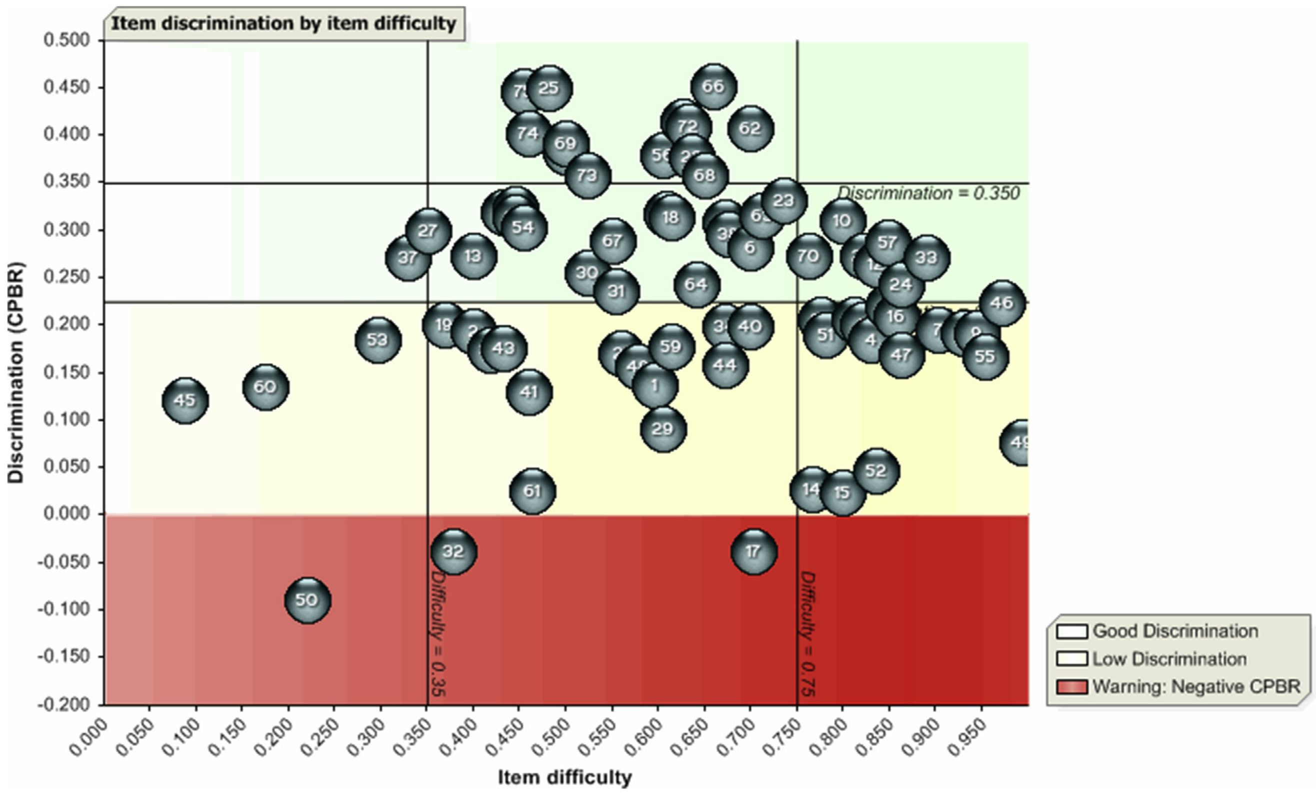

Figure 3: APS-III: Item discrimination/difficulty Year 2013–2014: (Case clusters-item 1–58, stand alone: item 59–75).

highest number of items) is shown in Table 8. The Myeloma PBL problem integrated anatomical pathology, chemical pathology, hematology, immunology, microbiology and pharmacology. The meningitis PBL problem integrated anatomical pathology, chemical pathology, immunology, microbiology, and pharmacology.

Discussion

The compilation of examinations requires more "effort" with case-based items.[20] The compilation of the APS-III examination paper was more challenging than that of APS-I and APS-II because the test items had to be well coordinated and the scenarios had to have a clear logical flow. Clinical scenarios progressively revealed information on the clinical presentation, complications, and laboratory and radiological investigations, and students in turn were assessed on their interpretation and management of specific conditions, in keeping with the views that test items should require multi-logical thinking,[21] which promotes critical thinking.[21] Answering items in case cluster (or context-rich MCQs) requires students to have basic information but also be able to apply it.[22] It would be difficult to pick out correct answers without properly analyzing and evaluating the clinical data as they are revealed.[22] Thus, it requires several Bloom's Taxonomy levels per case cluster,[22] which in itself is a "complex problem" that must be "holistically assessed".[22]

In this study, the number of items per scenario ranged from 2 to 8, depending on the topic, the sub-specialties involved, and what other questions were asked in the rest of the examination paper. (Some course content was examined in the EMQ section of the examination.) (Each sub-specialty had an approximately equal number of items in the examination paper.)

In integrated examinations, such as MEQs/PDQs, there may be an under- or over-representation of some sub-specialties.[6,9,18] This is also true of CS-MCQs. This speaks to the fact that some sub-specialties may not have relevant content and learning objectives for given scenarios. Furthermore, variation in numbers of items per scenario is unavoidable because it is necessary to ensure the examinations cover the depth and breadth of the syllabus in the integrated examination where the total number of MCQ items is 75. However, unlike MEQs/PDQs, more scenarios can be used,[8] which results in more content being examined. There were 18, 21 and 19 scenarios (Table 1) in the three years in this analysis. Assessment is also limited to "key issues",[8] another possible explanation as to why some scenarios had only 2 items. In addition, it has been shown that a higher reliability for tests occurred when patient cases used two to three test items, and the reliability and generalizability increased with an increased number of cases, not test items per case.[20] The case-based items increase the validity of examinations.[20] Clinical cases vary in length (and complexity); hence, the number of test items per case is unequal. This unbalanced design of items is also observed in the National Board Dental Hygiene examination.[20]

Reliability

Experts recommend high KR-20 reliability means. A high KR-20 result indicates a reliable test,[23] internally consistent instruments[24,25] and that the test is reproducible and consistent. A KR-20 value closer to 1 does a better job

Table 1: Analysis of integrated clinical scenarios vs. stand-alone MCQs in 3 years in APS-III.

	2011−2012			2012−2013			2013−2014		
No of students	202			194			221		
Number of clinical scenarios	18			21			19		
	Clinical scenarios	Stand alone	Total	Clinical scenarios	Stand alone	Total	Clinical scenarios	Stand alone	Total
	67 (89.3%)	8 (10.7%)	75 (100%)	63 (84.0%)	12 (16.0%)	75 (100%)	58 (73.3%)	17 (22.7%)	75 (100%)
1: Item difficulty (Pi)									
Mean (Range)	0.704 (0.059−0.980)	0.663 (0.168−0.901)	0.699 (0.059−0.980)	0.625 (0.134−0.990)	0.576 (0.103−0.912)	0.617 (0.103−0.969)	0.636 (0.090−0.955)	0.581 (0.176−0.851)	0.631 (0.090−0.955)
Number of items in 3 levels of difficulty (Pi)									
≥0.75	0	5	5	11	1	12	3	1	4
0.36−0.74	44	2	46	27	8	35	18	14	32
≤0.35	23	1	24	25	3	28	37	2	39
2: Item discrimination (Di) (corrected point biserial ratio − CPBR)									
CPBR mean (range)	0.234 (−0.116−0.454)	0.208 (0.043−0.329)	0.231 (−0.116−0.454)	0.188 (−0.092−0.394)	0.196 (0.034−0.380)	0.190 (−0.092−0.382)	0.211 (−0.092−0.412)	0.309 (0.024−0.449)	0.233 (−0.092−0.449)
Number of items in 5 levels of discrimination **(Di)** (CPBR)									
≥0.35	16	0	16	4	1	5	4	9	13
0.226−0.340	21	4	25	19	3	22	20	4	24
0.160−0.225	17	1	18	22	2	24	21	3	24
0.000−0.150	10	3	13	14	5	19	10	1	11
Negative	3	0	3	4	1	5	3	0	3
3: Bloom's taxonomy: number of items by level of Bloom's taxonomy									
Level I	6 (9%)	0	6 (8%)	0	2 (16.7%)	2 (2.7%)	3 (5.1%)	0	3 (4%)
Level II	6 (9%)	3 (37.5%)	9 (12%)	7 (11.1%)	1 (8.3%)	8 (10.7%)	9 (15.5%)	2 (11.8%)	11 (14.7%)
Level III	24 (35.8%)	0	24 (32%)	24 (38.1%)	4 (33.3%)	28 (37.3%)	21 (36.2%)	7 (41.2%)	28 (37.3%)
Level IV	31 (46.3%)	5 (62.5%)	36 (48%)	32 (50.8%)	5 (41.7%)	37 (49.3%)	25 (43.1%)	8 (47.1%)	33 (44%)

Table 2: Chi Square (χ^2) tests result with Yates' corrections for Table 1.

	2011—2012	2012—2013	2013—2014
χ^2 tests result for difficulty indices (3 levels)	34.83; $P < .01$	1.13; $P > .05$	13.07; $P < .01$
χ^2 tests result for discrimination levels (5 levels)	2.46; $P > .05$	1.66; $P > .05$	15.21; $P < .01$
χ^2 tests result Bloom's taxonomy levels (4 levels)	4.78; $P > .05$	5.26; $P > .05$	0.07; $P > .05$

of discriminating high performers from poorer performers. A KR-20 value of 0 shows no discrimination. This means the item is easy or a "confidence builder".[23] Less than 0.3 is a poor discriminator.[23] A negative KR-20 indicates an unreliable test.[24] A value of 0.7 is acceptable, and for longer examinations, e.g., with more than 50 items, a KR-20 value of 0.8 is desirable. Higher scores, >0.9, indicate that the examination is homogenous, which is a desirable characteristic. In this analysis, the test KR-20 means for APS-III were consistently high, indicating high reliability. The CS-MCQs had high KR-20 (examples in Table 8), consistently >0.8. There were no statistically significant differences in the KR-20 reliability coefficient (Table 5) across the years in all three courses. This shows that the MCQ items were consistent (CS-MCQs and stand-alone) throughout.

Item difficulty (Pi) and Bloom's taxonomy cognitive levels

Statistically, except for 2011—2012 and 2013—2014 for Pi (Table 2), there was no significant difference between the CS-MCQs and stand-alone MCQs in APS-III. Acceptable levels of Pi were achieved, with the majority of items falling in the moderate difficult category whether in the CS-MCQ or stand-alone MCQs. Writers recommend a wide range of difficulties in test items.[4,17] The Medical Council of Canada, 2010, recommends a range of 0.2—0.9[26] (or 20—90%). If Pi is close to 0.00 or 1.00, the item needs to be improved or discarded because it is not giving any information about differences among examinees' trait levels or abilities. Kartik A. Patel et al. in 2013[14] used a lower range of Pi. If Pi was <30% or >70% it was considered unacceptable and the MCQ needed modification. If Pi was between 30% and 70% the item was acceptable. Between 50% and 60% was considered optimum. That being said, some teachers like to have a few items that are easy "to make students feel good about themselves".[4] However, examiners should be careful not to compromise the quality of the test.[4] Some teachers actually define the number of items at different levels of difficulty. Edwardo Beckhoff (2000)[27] 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%".

Differences in Pi in this study may be because some item constructors were more advanced in item construction than

Table 3: Item distractor analysis of clinical scenarios vs. stand-alone MCQs in APS-III.

Year	2011—2012			2012—2013			2013—2014		
Number of students	202			194			221		
Number of clinical scenarios	18			21			19		
	Clinical scenarios	Stand alone	Total	Clinical scenarios	Stand alone	Total	Clinical scenarios	Stand alone	Total
Total number of items	67 (89.3%)	8 (10.7%)	75 (100%)	63 (84.0%)	12 (16.0%)	75 (100%)	58 (73.3%)	17 (22.7%)	75 (100%)
No of items with all-functioning distractors	24 (35.8%)	2 (25%)	26 (34.7%)	22 (34.9%)	6 (50%)	28 (37.3%)	29 (50.0%)	8 (47.1%)	37 (49.3%)
No of items with non-functioning distractors	43 (64.2%)	6 (75%)	49 (65.3%)	41 (65.1%)	6 (50%)	47 (62.7%)	29 (50.0%)	9 (52.9%)	38 (50.7%)
Total number of distractors	201	24	225	189	36	225	174	51	225
Total no of non-functioning distractors (<5%)	69 (34.3%)	10 (41.7%)	79 (35.1%)	63 (33.3%)	7 (19.4%)	70 (31.1%)	64 (36.8%))	9 (17.6%)	73 (32.4%)
χ^2 (with Yates corrections)	0.046; $P > .05$			0.441; $P > .05$			0.004; $P > .05$		

Table 4: Analysis of MCQs in years 2011−2012, 2012−2013, and 2013−2014 in APS-I, II and III.

	2011−2012			2012−2013			2013−2014		
Course	APS-1	APS-II	APS-III	APS-I	APS-II	APS-III	APS-I	APS-II	APS-III
Number of students	200	196	202	202	199	194	227	224	221
Number of items	75 (100%)	75 (100%)	75 (100%)	75 (100%)	75 (100%)	75 (100%)	75 (100%)	75 (100%)	75 (100%)
Mean	43.215	44.510	52.460	44.812	38.829	46.230	44.696	42.549	47.326
Median	44.000	45.000	53.500	45.000	39.000	47.000	45.000	42.000	47.000
Mode	42.000	55.000	61.000	50.000	41.000	52.000	46.000	37.000	47.000
Standard deviation	8.201	8.509	8.448	7.724	7.950	7.723	7.379	7.385	9.240
Variance	67.255	72.405	71.374	59.656	63.203	59.646	54.443	54.536	85.384
Max score	62	61	69	67	65	63	66	61	70
Min score	19	21	28	27	17	18	29	21	23
Standard error of mean	0.580	0.608	0.594	0.543	0.564	0.554	0.490	0.493	0.622
Standard error of measurement	3.786	3.667	3.467	3.736	3.846	3.629	3.618	3.736	3.678
KR-20-reliability	0.787	0.814	0.832	0.766	0.766	0.779	0.760	0.744	0.842
Spearman−Brown split half reliability coefficient	0.784	0.803	0.830	0.770	0.765	0.778	0.759	0.744	0.834
Spearman−Brown prophecy reliability formula	0.879	0.891	0.907	0.870	0.867	0.875	0.863	0.853	0.910
Guttman split-half reliability coefficient	0.782	0.802	0.829	0.770	0.763	0.777	0.756	0.742	0.831
Difficulty mean (range)	0.576 (0.070−0.965)	0.593 (0.010−0.908)	0.699 (0.059−0.980)	0.597 (0.064−0.960)	0.518 (0.111−0.874)	0.617 (0.103−0.969)	0.596 (0.026−0.974)	0.567 (0.022−0.924)	0.631 (0.090−0.955)
CPBR mean (range)	0.191 (−0.142−0.524)	0.211 (−0.114−0.500)	0.231 (−0.116−0.454)	0.180 (−0.051−0.407)	0.176 (−0.162−0.410)	0.190 (−0.092−0.382	0.171 (−0.080−0.342)	0.160 (−0.154−0.366)	0.233 (−0.92−0.449)

Table 5: Significant differences between reliability scores (KR-20) for 3 courses in 3 academic years.

Course	Academic years	Significant Differences between reliability scores (KR-20)
APS-1	2011–12 vs. 2012–13	z = 0.53, P > .05 not sig
	2011–12 vs. 2013–14	z = 0.69, P > .05 not sig
	2012–13 vs. 2013–14	z = 0.15, P > .05 not sig
APS-II	2011–12 vs. 2012–13	z = 1.26, P > .05 not sig
	2011–12 vs. 2013–14	z = 1.82, P > .05 not sig
	2012–13 vs. 2013–14	z = 0.52, P > .05 not sig
APS-III	2011–12 vs. 2012–13	z = 1.50, P > .05 not sig
	2011–12 vs. 2013–14	z = 0.34, P > .05 not sig
	2012–13 vs. 2013–14	z = 1.87, P > .05 not sig

others. However, because in this format, each question asks different aspects of a given case, it may also be more likely that for different sub-specialties, for a given topic, particularly in the CS-MCQs, the related objectives require different cognitive level skills by Bloom's taxonomy, as previously stated. One subspecialty may ask questions based on simple objectives, e.g., listing risk factors of a particular condition, and another may ask for more difficult aspects e.g., explaining the pathogenesis of conditions. One subspecialty may ask for the interpretation of specific results or problem solving, management and complications, which require higher level thinking, as shown in the example of Myeloma where the Pi values range from 0.353 to 0.864 and in the meningitis problem where the Pi values range from 0.299 to 0.953. The cognitive levels of Bloom's taxonomy in these two scenarios ranged from level II to level IV, with most falling in the Level III and IV groups. Higher taxonomy level questions encourage students to think deeper, learn better and retain more.[22] Case studies must be designed to require knowledge of multi-logical thinking.[32]

In comparison, the mean total scores for the classes, and the maximum scores are higher each year in APS-III than in APS-I and APS-II, yet the Pi means are also higher in APS-III than in APS-I and APS-II (Table 4). A possible explanation may be that the APS-III course is in semester 2. Students may be more comfortable with examinations in semester 2. Furthermore, all students would have rotated

Table 6: Distractor analysis of APS-I, II and III over three years.

	APS-I			APS-II			APS-III		
	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	75	75	75	75	75	75	67	63	58
No of items with all-functioning distractors	35 (46.7%)	35 (46.7%)	36 (48%)	32 (42.7%)	49 (65.3%)	34 (45.3%)	24 (35.8%)	22 (34.9%)	29 (50.0%)
No of items with non-functioning distractors	40 (53.3%)	40 (53.3%)	39 (52%)	43 (57.3%)	26 (34.7%)	41 (54.7%)	43 (64.2%)	41 (65.1%)	29 (50.0%)
Total number of distractors	225	225	225	225	225	225	201	189	174
Total no of non-functioning distractors (<5%)	53 (23.6%)	59 (26.2%)	70 (31.1%)	62 (27.6%)	32 (14.2%)	64 (28.4%)	69 (34.3%)	63 (33.3%)	64 (36.8%)
χ^2 (with Yates corrections	0.036; P > .05			9.213; P < .01			3.584; P > .05		

Table 7: Sub-specialty total score Pearson correlation coefficients: APS-III in 2013–2014.

	Anatomical pathology	Chemical pathology	Hematology	Immunology	Microbiology	Pharmacology
Anatomical pathology	1					
Chemical pathology	0.345 (P = 1.378E-007)	1				
Hematology	0.334 (P = 3.754E-007)	0.369 (P = 1.565E-008)	1			
Immunology	0.317 (P = 1.547E-006)	0.413 (P = 1.689E-010)	0.465 (P = 3.004E-013)	1		
Microbiology	0.306 (P = 3.703E-006)	0.363 (P = 2.839E-008)	0.244 (P = 2.473E-004)	0.271 (P = 4.363E-005)	1	
Pharmacology	0.476 (P = 6.539E-014)	0.468 (P = 2.052E-013)	0.461 (P = 4.999E-013)	0.392 (P = 1.533E-009)	0.376 (P = 7.996E-009)	1

Table 8: Example of analysis of 2 integrated clinical scenario MCQs.

Integrated clinical scenario: Multiple myeloma: Year 2013—2014 (8 items)					Integrated clinical scenario: Meningitis: Year 2012—2013 (5 items)			
Item	Specialty	Pi (Total test mean − 0.631)	Di (Total test mean − 0.233)	KR-20 (Total test mean − 0.842)	Specialty	Pi	Di	KR-20
Q1	Hematology	0.864	0.241	0.840	Microbiology	0.783	0.188	0.841
Q2	Immunology	0.484	0.447	0.836	Chemical pathology	0.837	0.045	0.842
Q3	Hematology	0.824	0.271	0.839	Immunology	0.299	0.182	0.841
Q4	Hematology	0.353	0.297	0.839	Microbiology	0.457	0.300	0.839
Q5	Anatomical pathology	0.448	0.321	0.838	Anatomical pathology	0.955	0.165	0.841
Q6	Hematology	0.606	0.089	0.842				
Q7	Hematology	0.525	0.253	0.839				
Q8	Pharmacology	0.557	0.234	0.840				

through all clerkships at this stage. In clerkships, students are in smaller groups, have closer contact with lecturers and receive clinical and practical application of all basic knowledge. They see more relevance in their studies and hence learn more as they gain better understanding of their subjects.[28] It could also be suggested that the clinical scenarios improve question clarity and increase relevance to the curriculum, as seen in the literature.[11] In the CS-MCQs, there were no "cued" items. However, in 2011—2013 and 2012—2013, two case clusters were hinged each year, and five were hinged in 2013—2014. Hinging does make the examination difficult.[12] In the study by Tsai et al.,[20] the case-based items were more difficult.

Item discrimination (Di)

In APS-III, statistically, except for 2013—2014 for Di (Table 2), there was no significant difference between the CS-MCQs and stand-alone MCQs. The moderately and highly difficult items showed higher discrimination than the easier items, similar to other reports.[4] In comparison, the Di means each year in APS-III were higher than those in APS-I and APS-II (Table 4). Staff members were informed of the items with poor Di (with negative CPBR) and were advised to modify them in their question banks. These items were also removed from the students' final examination results. For MCQ discriminators, most writers recommend a discrimination coefficient of $\geq$0.20.[4] Some may go as low as 0.15 and others as high as 0.25. DiBattista et al.[4] showed a curvilinear relationship between Pi and Di, which had been shown by others before. They also found that the tests with lower mean discrimination coefficients also had the lowest adjusted values of Cronbach's alpha. James Ware et al. (2008)[13] created arbitrary levels of discrimination power, where $>$0.4 was excellent, 0.30—0.39 was good, 0.15—0.29 was moderate and below 0.15 was considered to have no discrimination power of significance. They showed that over four years, the excellent category ranged from 0.8% to 21% and the very good category ranged from 10 to 19%. In a 2013 study by Kartik A. Patel et al.,[14] Di $<$ 0.20 was considered to be unacceptable, and the corresponding MCQ item required modification. Di values of 0.20—0.24 were acceptable and DI values of 0.25—0.34 were considered good. If Di $=$ 0.35 or more, they considered this as excellent discrimination. In their study involving 151 students taking a 50 MCQ test, the analysis showed 9 MCQ items to have a Di a value of 0.20, 5 items to have a Di value of 0.2.0—0.24, 16 items to have Di value of 0.25—0.34 and 20 items to have a Di value of $\geq$0.35. Most items fell into the acceptable difficulty range. To promote enhanced critical thinking, test items need to have a high level of discriminatory power.[21]

Item distractors

The power of discrimination is very dependent on the distractor options in the items. The discriminating power increases as the number of functioning distractors increases.[4] In APS-III overall, there was a high percentage of functioning distractors in the CS-MCQs (66.7—63.7%) and stand-alone MCQs (85.8—63.2%). There were no statistically significant differences regarding the number of items with all-functioning versus non-functioning distractors in the CS-MCQs. Comparing APS-I, APS-II and APS-III, the number of functioning distractors over the three years was still high (Table 6). There were no statistically significant differences across the years in APS-I and APS-III with regards to the number of items with all-functioning distractors and non-functioning distractors. For APS-II, there were statistically significant differences across the years.

Marie Tarrant et al. (2009),[29] in their study among nursing students, showed that only 13.8% items had functioning distractors of $>$5% in 4 or 5 option MCQs, stating that it is difficult to construct plausible distractors for most teachers. Most distractors really are just "fillers". They emphasized that the key is really the quality of the distractor and not so much the number of distractors, even suggesting reducing the options to just three. However, some researchers argue that the reduction to 3 options increases the chances of weaker students just guessing the correct answers. Increasing the number of distractors decreases the probability of guessing.[30] More options are associated with increased reliability and validity.[31] However, increasing the number of options increases the test time.[31] Furthermore, high-quality, well-constructed distractors reduce issues associated with cueing.[22]

Limitations

- The number of stand-alone MCQs is much small than the CS-MCQs for a fair comparison in APS-III. (Hence the analysis of APS-I and APS-II was used for comparison). Whereas APS-III examines different content, the same students take APS-I and APS-II and there are equal numbers of items in all courses. The same staff that taught and examined APS-III taught and examined APS-I and APS-II. All examination papers and answer keys were reviewed (for content, accuracy, cues and flaws) and approved by the examinations core committee and the head of the department. Other authors recommend this vetting of items by an interdisciplinary team[32] to ensure proper content, good quality and acceptable difficulty. Furthermore, all examination papers were reviewed by an external examiner, prior to the students taking them. Tsai et al.,[20] in their analysis, which had fewer case-based items than discipline-based items, showed that the reliability (Cronbach's alpha) was lower in case-based items compared to discipline-based items.
- The Pearson correlation analysis[33] between sub-specialties was performed on the whole examination and not separated into CS-MCQs and stand-alone MCQs. A question may be raised that CS-MCQs have higher correlations compared to stand-alone MCQs and hence the internal consistency reliability may tend to be hyper-inflated. In an earlier study, the authors[34] showed that the correlations between different sub-specialties (in the same department of Para-clinical Sciences) were strong among multiple modes of assessment: PDQ, MCQ and EMQ.[34] Inter-case correlations were not performed.
- This study did not document the views of the students on CS-MCQs. In a study in Ireland among marketing students, Christina Donnelly (2014)[35] reported that more than 50% students said that CS-MCQs were more difficult and more challenging because they made them think more and apply knowledge to the situations. They said that it took them longer to read and process the case study and hence answer the items. The other 50% of students suggested it was 'easier' and found that the case studies helped to stimulate answers, apply their learning from the lectures and use more of their own interpretation.
- This study did not document the views of staff on the CS-MCQs either. However, the compilation of the APS-III examination paper was more time consuming and challenging as stated earlier. Donnely's team reported that the lecturers found that the introduction of CS-MCQs provided a higher level of learning and more critical thinking for students and helped students blend theory with practice. They also commented that this assessment would be more time intensive for instructors to create. This study did not analyze the time taken for the individual CS-MCQs as was performed by Hays et al. in 2009.[11] However, APS-I, II, and III are all three hours long, and students did not report running out of time.

Conclusions

The goal of PBL is to integrate basic sciences and clinical specialties, helping students to learn better and improving their clinical reasoning.[36] Integrated CS-MCQs compare favorably to stand-alone MCQs. They are easy to align with PBL learning objectives, reliable and practical. They reflect and demonstrate effective learning and understanding, requiring students to think deeper for longer.[35] Focusing on key features allows a wider range of cases.[37] CS-MCQs can be constructed with rigorous psychometric standards to distinguish high and low scorers.[38] A high number of non-functioning distractors decrease the distractor efficiency and make items easier.[39] An item-analysis data review is recommended to improve MCQ items.[13]

Recommendation: The continued use of CS-MCQ is recommended.

Abbreviations: APS-I, Applied Para-clinical Sciences-I; APS-II, Applied Para-clinical Sciences-II; APS-III, Applied Para-clinical Sciences-III; CA, continuous assessment; CS, clinical scenario; CS-MCQ, clinical scenario multiple choice question; CPBR, corrected point-biserial ratio; EMQ, extended matching questions; KR-20, KuderRichardson-20; MCQ, multiple choice questions; MEQ, modified essay questions; PBL, problem based learning; PDQ, progressive disclosure questions; SAQ, short answer questions.

Conflicts of interest

The authors have no conflict of interest to declare.

Grants/Funding

None.

Ethical approval

Obtained from the Ethics Committee and office of The Dean, Faculty of Medical Sciences, University of The West Indies, St Augustine, Trinidad and Tobago.

Authors' contributions

Both authors contributed in this paper. SV conceived and designed the study, collected and analyzed the data, drafted the manuscript, reviewed and approved the final draft, reviewed and approved the corrections and submitted the initial and revised manuscripts. BS conceived and designed the study, collected and analyzed the data, drafted the manuscript, reviewed and approved the final draft, performed the final reference and Turnitin checks and reviewed and approved the corrections.

Acknowledgments

The authors would like to thank (1) Prof Samuel Ramsewak, former dean; (2) the (former) head and staff of the department Para-clinical Sciences; and (3) Mrs Louise Green, former staff of the Assessment Unit, Centre for Medical Sciences Education (CMSE), – Faculty of Medical Sciences.

References

1. Tabish SA. Assessment methods in medical education. **Int J Health Sci (Qassim) 2008**; 2(2): 3−7.
2. Azer SA. Assessment in problem-based learning. **Biochem Mol Biol Educ 2003**; 31(6): 428−434.
3. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. **BMC Med Educ 2007**; 7: 49. 10.1186/1472-6920-7-49, http://www.biomedcentral.com/1472-6920/7/49.
4. DiBattista D, Kurzawa L. Examination of the quality of multiple-choice items on classroom tests. **Can J Scholarsh Teach Learn 2011**; 2(2). Article 4. Available at: http://ir.lib.uwo.ca/cjsotl_rcacea/vol2/iss2/4.
5. von Bergmann H, Dalrymple KR, Wong S, Shuler CF. Investigating the relationship between PBL process grades and content acquisition performance in a PBL dental program. **J Dent Educ September 2007**; 71(9): 1160−1170.
6. Moeen-uz-Zafar-Khan, Aljarallah BM. Evaluation of modified essay questions (MEQ) and multiple choice questions (MCQ) as a tool for assessing the cognitive skills of undergraduate medical students. **Int J Health Sci (Qassim) 2011**; 5(1).
7. American Educational Research Association, American Psychological Association, National Council of Measurement in Education. *Standards for educational and psychological testing*. Washington, DC: AERA; 1999.
8. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. **Lancet 2001**; 357: 945−949.
9. Palmer EJ, Duggan P, Devitt PG, Russell R. The modified essay question: its exit from the exit examination? **Med Teach 2010**; 32: 300−307.
10. Campbell DE. How to write good multiple-choice questions. **J Paediatr Child Health 2011**; 47: 322−325.
11. Hays RB, Coventry P, Wilcock D, Hartley K. Short and long multiple-choice question stems in a primary care oriented undergraduate medical curriculum. **Educ Prim Care 2009**; 20(3): 173−177.
12. Case SM, Swanson DB. *Constructing written test questions for the basic and clinical sciences*. 3rd ed. (Revised) 2002. National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104.
13. 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): 238−243. http://dx.doi.org/10.1080/01421590802155597.
14. Patel KA, Mahajan NR. Itemized analysis of questions of multiple choice question (MCQ) exam. **Int J Sci Res 2013**; 2(2): 279−280.
15. Anastasi A, Urbina S. *Psychological testing*. Delhi: Pearson Education, Inc; 1997.
16. Linn RL, Gronlund NE. *Measurement and assessment in teaching*. Pearson Education, Inc; 2000.
17. Ebel RL, Frisbie DA. *Essentials of educational measurement*. New Jersey: Prentice Hall, Inc; 1991.
18. Vuma S, Sa B. Evaluation of the effectiveness of progressive disclosure questions as an assessment tool for knowledge and skills in a problem based learning setting among third year medical students at The University of The West Indies, Trinidad and Tobago. **BMC Res Notes 2015**; 8(673). http://dx.doi.org/10.1186/s13104-015-1603-0. Available at: http://www.biomedcentral.com/content/pdf/s13104-015-1603-0.pdf [Accessed 14 November 2015].
19. Huitt W. *Bloom et al.'s taxonomy of the cognitive domain. Educational Psychology Interactive*. Valdosta, GA: Valdosta State University; 2011. Available at: http://www.colorado.edu/AmStudies/lewis/1025/bloomtax.pdf [Accessed 21 August 2015].
20. Tsai T-H, Dixon BL, Littlefield JH. Constructing licensure exams: a reliability study of case-based questions on the national board dental hygiene examination. **J Dent Educ 2013**; 77(12): 1588−1592.
21. Morrison S, Walsh K. Writing multiple-choice test items that promote and measure critical thinking. **J Nurs Educ 2001**; 40(1). ProQuest Central.
22. 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. http://www.biomedcentral.com/1472-6920/14/249.
23. KR-(20). Available at: http://eacvisualdata.com/eacs/kr20.aspx. [Accessed 26 September 2015].
24. Thompson NA. KR-20. Available at: http://knowledge.sagepub.com/view/researchdesign/n205.xml. [Accessed 21 August 2015].
25. Kuder and Richardson Formula 20. Available at: http://www.real-statistics.com/reliability/kuder-richardson-formula-20/. [Accessed 26 September 2015].
26. Medical Council of Canada. *Guidelines for the development of multiple-choice questions*. Available at: http://mcc.ca/wp-content/uploads/Multiple-choice-question-guidelines.pdf; February 2010 [Accessed 30 July 2015].
27. Backhoff E, Larrazolo N, Rosas M. The level of difficulty and discrimination power of the basic knowledge and skills examination (EXHCOBA). **Rev Electrón Investig Educ 2000**; 2(1). Available at: http://redie.uabc.mx/vol2no1/contents-backhoff.html [Accessed 14 November 2015].
28. 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. **Caribb Teach Sch 2015**; 5(1): 25−35.
29. 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. http://dx.doi.org/10.1186/1472-6920-9-40. Available at: http://www.biomedcentral.com/1472-6920/9/40 [Accessed 14 November 2015].
30. Considine J, Botti M. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. **Collegian 2015**; 12: 1.
31. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice Item-writing rules. **Appl Meas Educ 1989**; 2(1): 51−78. Published online: 14 Dec 2009.
32. Johnson TR, Khalil MK, Peppler RD, Davey DD, Kibble JD. Use of NMBE comprehensive basic science examination as a progress test in the preclerkship curriculum of a new medical school. **Adv Physiol Educ 2014**; 38: 315−320. http://dx.doi.org/10.1152/advan.00047.2014. http://advan.physiology.org.
33. Pearson's r Correlation: Available at: http://faculty.quinnipiac.edu/libarts/posci/Statistics.html. [Accessed 26 September 2015].
34. 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. **Caribb Teach Sch 2015**; 5(1): 37−46.
35. Donnelly C. The use of case based multiple choice questions for assessing large group teaching: implications on student's learning. **Ir J Acad Pract 2014**; 3(1). Article 12. Available at: http://arrow.dit.ie/ijap/vol3/iss1/12 [Accessed 14 November 2015].
36. Ehmelo CH, Gotterer GS, Bransford JD. A theory-driven approach to assessing the cognitive effects of PBL. **Instr Sci 1997**; 25: 387−408.
37. Bordage G. *Assessing clinical decision making: focusing only on the critical, challenging decisions, the key features*. Available at:

http://cores33webs.mede.uic.edu/dmefac/bordage/presentations/KeyFeatures_Chiba_0307.pdf; 2007 [Accessed 18 August 2016].

38. ACGME. *Toolbox of assessment methods*. ACGME outcomes project, Accreditation Council for Graduate Medical Education, American Board of Medical Specialties. Available at:, http://njms.rutgers.edu/culweb/medical/documents/ToolboxofAssessmentMethods.pdf; 2000 [Accessed 18 August 2016].

39. Gajjar S, Sharma R, Kumar P, Rama M. Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad. **Indian J Community Med 2014**; 39(1): 17—20.

How to cite this article: 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 Sc 2017;12(1):14—26.