Quality of Multiple Choice Questions in Fixed Prosthodontic Module

1st Author
Ehab Mohamed Abdelhlim Ahmed
Department of Conservative Dentistry
University of Gezira
Gezira, Sudan

1st Author
Nizar Ismail
Trauma and Orthopaedics Registrar
Royal Cornwall Hospital
Truro, United Kingdom

2nd Author
Klaudyna Ziolkowska
General Surgery Speciality Trainee
Royal Glamorgan Hospital
Pontyclun, United Kingdom

3rd Author
Israa Ahmed Ali
Sudan Ministry of Health
Khartoum, Sudan

4th Author
Elhadi Mohieldin Awooda
Education and Development Center
University of Gezira
Gezira, Sudan

5th Last Author
Magda Elhadi Ahmed Yousif
Professor of Community Medicine
University of Gezira
Gezira, Sudan

Abstract:- The objective of this study aimed to evaluate the quality of Multiple Choice Questions (MCQs) in Fixed Prosthodontic Department in Faculty of Dentistry in university of Gezira. This evaluation study was conducted over a period of four year exam, from 2014 to 2017 and evaluated four final semester assessment examinations comprising of MCQs cognitive level using Bloom Taxonomy and Item Writing Flaws (IWFs) employing a validated check list based on the guidelines of the NBME (National Board of Medical Examinations). A total of 80 MCQs were analyzed. Items were classed as flawed if they contained one or more than one flaw. The result of study found about half of the MCQs - 45 (56.3%) were assessing the recall of information, while 17(21.3%) were assessing the application of knowledge, 9 (11.3%) were assessing comprehension, 5 (6.3%) and 4 (5%) were assessing interpretation of data. The total of flawed items out of 80 items in four exams was 132 flaws. Most common types of flaws were grammatical mistakes and hand cover test pass 42% (32) and negatively constructed items 29%. The study concludes that cognitive level of assessment tools MCQs is low, and IWFS are common in the MCQs. Therefore, educators should be encouraged and trained to design problem-solving questions which are devoid of flaws.

Keywords:- *MCQs; IWFs; NBME; medical examinations; Bloom Taxonomy; dentistry.*

I. INTRODUCTION

1.1 Background

Well-constructed cognitive assessment tools encourage the students to engage their higher cognitive skills and abstract thinking. There are several ways to assess the knowledge domain including Free response examinations (Long Essay Questions, Short answer Questions, Modified Essay questions), Multiple choice questions, Key feature questions, peer-assessment and self- assessment. Each of those has its pros and cons and assesses different levels of bloom's taxonomy. No single method of evaluation is superior to other and probably a valid and reliable evaluation requires a mixture these method [6].

In addition to thinking about the topics that are important to include on a test, you should think about how to structure those questions to test more than the recall isolated facts. Classically, test questions have been described as requiring recall, interpretation, or problem solving (memory, comprehension, and reasoning) depending on the cognitive processes required to answer the question. Typical definitions refer to “Recall Questions” as those which evaluate the student’s knowledge of definitions or isolated facts. “Interpretation Questions” require student to review some information, often in tabular or graphic form, and reach some conclusion (e.g., a diagnosis). “Problem-Solving Questions” present a situation and require test-taker to take some action (e.g., the next step in patient management) [17]. The challenge with these classifications is that the cognitive processes required to answer a question are as dependent on the question content as wells as on the background of the examinee. Experts in a content area may simply recall an answer with little or no conscious effort, whereas others may need to deduce the answer from basic principles. The cognitive processes involved in responding to a question are examinee-specific, making the taxonomic approach difficult to use [8].

1.2. Problem identification

Recently a large number of schools in the medical field have been opened, most of the expert teachers have moved outside the country, leaving only a few numbers of teachers to cover these medical schools. This has led to a shortage in the teaching system. Thus, to cover it teachers started to teach in several universities, which could potentially affect the quality of assessment. ‘Health Professions Education’ is new in this country, with only two universities - Gezira university Khartoum university currently providing post graduate studies in this field. This led to a shortage in training medical teachers in fields like assessment and constructing MCQs, which leaves most assessments to be based on the experience of medical teachers although it is largely proven that faculty members who received formal training in constructing MCQs can write better items than those who are only expert in teaching [5].

Beside of the lack of proper training on the construction of MCQs, another fact to be considered is the extreme reuse of the MCQs items. Although it is a usual trend to maintaining, but it's required to regularly renew the item [5].

1.3 Justification

Students' learning is largely enhanced by assessment, thus development of high quality tests is an important skill for educators [4].

MCQs can be time consuming and difficult to produce even for those educators who have received formal training. Well-made MCQs result in impartial assessment of the scholar and can measure knowledge as well as comprehension, application and analysis [5].

Poorly constructed MCQs can lead to unreliable and invalid results, which is reflected on the whole educational system - this research will (may) be the first step to determine our situation in order to improve it [1].

1.4 Objectives

1.4.1 General objectives

To evaluate the quality of multiple choice questions of Fixed Prosthodontic module, Faculty of Dentistry, Gezira University (2014-2017).

1.4.2 Specific objectives

i. To determine the different cognitive levels (according to blooms taxonomy) used in the items
ii. To determine the conformity of the items writing with the given MCQS items writing guide line [18]

II. LITERATURE REVIEW

Students' learning is largely driven and enhanced by assessment, thus development of high quality tests is an important skill for educators. The mode of assessment has been shown to affect the students' learning process. Usually, educators develop the test items by themselves or sometimes rely on item test banks as a source of questions [3] [10].

MCQs are adequate competency tests for assessing knowledge and comprehension that can be designed to measure application and analysis. Use of well-designed MCQs has been increasing significantly due to their higher reliability, validity and ease of scoring. Also, well-constructed MCQs can effectively test higher levels of cognitive reasoning and can help differentiate between high and low-achieving students. Despite the above various studies have documented violation of MCQs' construction guidelines. Single best questions (SBAs) and MCQs or are time-consuming and difficult to construct, even for educators who are formally trained on this matter. Well-constructed MCQs promote impartial testing and have the potential to measure knowledge, comprehension, application and analysis [3] [7].

A typical MCQ item consists of a question (referred to as the stem) and a set of two or more options that consist of possible answers to the question. The student's task is to choose the one option that provides the best answer to the question posed. The best answer is referred to as the keyed option and the remaining options are called distracters. For teachers, an important benefit of using MCQ items in classroom tests is that marking tends to be easy and quick, especially when students put down their answers on an optically scanned response sheet, such as the commonly used Scantron® form [2] [7] [8]. Ease of marking can make MCQ testing particularly appealing to instructors who look after large cohorts of students. Another significant benefit is that a well-constructed MCQ examination can yield test scores at least as reliable as those obtained by a constructed-response test whilst simultaneously allowing for coverage of a wide spectrum of topics [12].

Another approach, perhaps more simple and objective, bases item classification on the task asked of the test taker. If an item requires an examinee to make a prediction, reach a conclusion or select a course of action or, it should be classed as an application of knowledge item. If an item only tests only recollection of isolated facts (without requiring their application), it should be classified as a recall item. All items should require application of knowledge, allowing assessment of both an examinee's information base plus ability to use that information [2].

To produce effective MCQs, one must strive to produce items that are free from flaws. Item writing flaws in MCQs can affect students' performance by resulting in items that are either more or less difficult to be answered. It is found by certain authors that flaws typically make the item less difficult. Additionally, flawed items may yield a certain ambiguity, without which 25% of failed students would have passed the exams [11].

Item-writing flaws (IWFs) arise when we veer from the accepted MCQ writing recommendations. As a consequence, such MCQs affect the performance of the students making it either more difficult or easier for them to answer the questions [14].

Foundation for Developing and Validating Test Items, It investigates the role of validity in test item development, processes of item development as well as advantages and limitation of cognitive taxonomy. The authors explained the concept of test item referring to other worthwhile work, and the planning and development process of both selected response (SR) items and constructed response (CR) items [16]. They also discussed the recruitment of item writers and their main tasks, training and guidelines on item writing, and reviewing test items looking at factors such as fairness and complexity of language. With regard to cognitive demand as well as content of test items, it describes the drawbacks of the cognitive taxonomy for classifying cognitive demand and also adds recommendations in the context of knowledge, skills and abilities. The section analyses the concept of cognitive ability being represented by a model of learning which includes: a) specification of declarative and procedural knowledge, b) a measurement plan, c) hypotheses and evidence that accepts or rejects hypotheses, d) description of threads that move learners from novice toward expert, e) consideration of factors affecting learning, and f) consideration of construct irrelevant variance that may diminish validity. Further the discussion on procedure of item format generates four fundamental types of item formats i.e. subjective versus objective scoring, production versus selection, free-response versus fixed-response, and performance versus product. Each format has limitations and advantages for the users in view of the nature of the subject and objectives to be measured [9].

Guidelines for Writing MCQs

1. Each MCQ should assess an important theme or sub theme.
2. MCQs should test more than just recall of facts.
3. MCQs ought to assess higher levels of cognition i.e. interpretation and application of knowledge, analysis of data, critical thinking and problem solving.
4. Framework of a MCQ:

a) stem/clinical scenario consisting of a vignette linked to a theme. Simple language should be used in order to be easy to understand by students. The stem can include information such as:

- age, gender (e.g. a 80-year-old man)
- location (e.g. emergency department)
- presenting complaint (e.g. pain)
- site (e.g. chest / abdomen)
- duration (e.g. minutes/hours/days)
- past medical history (e.g. heart failure)
- physical findings (e.g. murmur)
- investigations (significant postive or negative findings)
- management (operative/medical, can include treatment and response to it)

b) lead in that poses a clear question in relation to the stem. It is the main task and should be corresponding to the sub-theme.

c) options: four or more options. One of them is the most appropraite i.e. best answer (correct answer / key) while the remaining options serve as plausible distracters. All options should be of similar length and reasonably brief.

5. Avoid:

- use of double negatives in stems such as "the treatment of this condition does not consist of the following meassures except."
- Using the word "except."
- Options 'all of the above' and 'none of the above', as these increase the chance of guessing.
- Non-specific phrazes such as 'rarely' and 'usually', absolute terms such as 'always' and 'never'.
- Grammatical errors, inconsistencies or cues.
- Long correct answers – i.e. the correct answer is more detailed and extensive than other options
- Word repeats – the same phrase can be found in the stem as well as in the correct answer
- Superfluous information.
- Tricky and overly complex items [9] [18].

III. MATERIAL AND METHODS

3.1 Study design

Evaluation study.

3.2 Study area

The study was conducted at the Faculty of Dentistry, University of Gezira in Sudan established in 2001 as the first Dental School outside the capital Khartoum at that time. The philosophy of University of Gezira is serving the population by addressing community problems and strengthening preventive and therapeutic oral health services.

3.3 Study population

MCQs of the Previous, four Fixed Prosthodontic module final semester exams

3.4 Duration

Fixed Prosthodontic module final semester exams held between 2014-2017.

3.5 Sample size

Total coverage.

3.6 Study variables

Level of cognitive domains assessed
Academic years
Present of item flaws.

3.7 Data collection tools and technique

3.7.1 Data collection tools

A validated check list based on the guidelines of the NBME (National Board of Medical Examinations) was used for the evaluation.

3.7.2 Data collection technique

This analytical study was conducted after the completion of the assessments for the year 2017. The original MCQs that were submitted to the assessment committee for the purpose of summative exams were grouped according to the year then analyzed for assessment.

The exam for the years 2014, 2015, 2016 and 2017 in fixed prosthodontic subject was collected and reviewed by two experts. In each exam, there were 20 MCQs. A total of 80 MCQs was evaluated for their cognitive levels and item writing flaws (IWFs). The questions' cognitive levels were evaluated using the Bloom's taxonomy (1956). For identifying types of IWFs' standard criteria given by several educationists were used and frequently occurring violations of item-writing guidelines were selected from literature and were subsequently applied to assess the quality of the 80

MCQs in all four exams. The structure of each question was analysed for technical accuracy. Items were classified as ‘flawed’ if they contained one of the flaws. Frequently observed flaws were grouped into:

- Hand cover test pass
- Grammatical Mistakes
- Length of the correct answer
- Repeated word in the stem and options
- Merging more than one item in one answer
- Long and exhausting options
- Using absolute terms
- Using non-logical option
- Poorly arranged numeric data
- Negatively constructed items

- Usage of the “None of the above” or “all of the above” types of question
- Unclear or vague lead-in or options.

3.8 Inclusion criteria

All MCQS submitted for semester 10 fixed prosthodontic module, years 2014, 2015, 2016&2017.

3.9 Exclusion criteria

- Mid-term exams and supplementary exams
- Substitute exams and any final exam out of the period and subject of the study.

3.10 Data management and statistical analysis

Data was checked for completeness, consistency and range. Total numbers of items reviewed were calculated. Percentages of the technical flaws encountered were calculated with measurement of frequencies in each question. Chi-square test was used to analyze the improvement in categories of variables between the years. The data of each item was analyzed using SPSS 23. Results were displayed in appropriate tables constructed with Microsoft Office word.

3.11 Ethical consideration

Ethical approval was obtained from the ethical committee in faculty Dentistry University of Gezira. No ethical hazards have been found in the processing of this study.

IV. RESULTS AND DISCUSSION

4.1 Results

As shown in the tables 1 and 2 (see page 6), that demonstrates the quality of items found in MCQS material in various years, there is clear and stable improvement in the use of repeated word. Some improvement can be noticed in the hand cover test, grammatical mistakes and using non-logical structures. The remaining items did not follow a continuous sequence. However, merging more than one item in one answer and long and exhausting options improved in 2017. On the contrary - using absolute, poorly arranged numeric, negatively constructed items, using phrases: none of the above and all of the error increased in 2017. When considering 'very long correct' the results were stable throughout the four academic years. When comparing between different academic years in regards with the check list component of qualified MCQs the only significance values were grammatical mistake (P-value 0.036), negatively constructed Items (P-value 0.01) and using phrases: none of the above and all of the above (P-value 0.028).

In table 3 the total error percent of the four academic years is in a close range to each other, with the lowest percent 30 in 2016 and the highest 36 in 2017. However, there was regular improvement in MCQ material from the year 2014, 2015 and 2016 with 34, 32 and 30%, respectively. Unfortunately, this progress declined in 2017 when the error rate increased to 36%. Overall, approximately two third of the MCQ material written were free of errors.

Table (1): Demonstrate items found in MCQs material in various years and their p-value

Check List Items	***2014***		***2015***		***2016***		***2017***		***2 sided P-value***
	Yes	***No***	***Yes***	***No***	***Yes***	***No***	***Yes***	***No***	
Hand cover test	13	7	15	5	10	10	8	12	**.115**
Grammatical mistake	10	10	13	7	6	14	5	15	.039
Very long correct	0	20	0	20	1	20	0	20	.539
Repeated word	3	17	1	19	1	19	1	19	.539
Merging more than one item in one answer	2	18	3	17	1	19	1	19	.632
Long and exhausting	0	20	0	20	1	19	0	20	.386

Table (2); Demonstrate items found in MCQs material in various years and their p-value

Check List Items	***2014***		***2015***		***2016***		***2017***		
	Yes	***No***	***Yes***	***No***	***Yes***	***No***	***Yes***	***No***	***P-value***
Using absolute	1	19	0	20	0	20	2	18	.200
Using non-logical	4	16	1	19	1	19	0	20	.081
Poorly arranged numeric	0	20	1	19	0	20	1	19	.420
Negatively constructed Items	3	17	2	18	8	12	10	10	.010
Using phrases : none of the above and all of the above	1	19	5	15	0	20	1	19	.028
Unclear OR vague lead-in or option	3	17	2	18	2	18	3	17	.928

Table (3): displaying the percentage of MCQs material errors during the four years calculated from table 1 and 2	
Academic year	***Error percent (%)***
2014	34%
2015	32%
2016	30%
2017	36 %

When looking at each of the check list component separately in table 4, it can be concluded that the highest percent of error is shared by the hand cover test and grammatical mistakes, both at 42%. Followed by Negatively constructed Items with (29%). Whereas, the remaining component lay in the range between (0% to 13%).

Table (4): error percent in check list components calculated from table 1and 2	
Check List	***Result in percent (%)***
Hand cover test	42%
Grammatical mistake	42%
Very long correct	0%
Repeated word	7%
Merging more than one item in one answer	9%
Long and exhausting	1%
Using absolute	4%
Using non-logical	7%
Poorly arranged numeric	3%
Negatively constructed Items	29%
Using phrases : none of the above and all of the above	9%
Unclear OR vague lead-in or option	13%

As noted in table 5 below, the majority of questions were based on knowledge (56.3%), 21.3% on application, comprehension-based questions were at 11.3% and 6.3% assessed analysis.

Table (5): Demonstrating the association between Academic Year and Assessed Cognitive Domain according to number of questions					
Assessed Cognitive Domain	***Academic Year***				
	2014	***2015***	***2016***	***2017***	***Total***
Knowledge (recall)	10	13	9	13	45
	50.0%	65.0%	45.0%	65.0%	56.3%
Comprehension	0	2	6	1	9
	0.0%	10.0%	30.0%	5.0%	11.3%
Application	3	4	5	5	17
	15.0%	20.0%	25.0%	25.0%	21.3%
Analysis	5	0	0	0	5
	25.0%	0.0%	0.0%	0.0%	6.3%
Evaluation	2	1	0	1	4
	10.0%	5.0%	0.0%	5.0%	5.0%
Pearson Chi-Square value [28.003], 2-sided P-value [0.006]					

The Chi-Square Test value were found to be (28.003) with significant P-value (0.006).

In 2014, 65% of the hand cover test passed. In 2015, 75% were passed which was the highest percentage. In 2016 the percentage decreased to 50%. In 2017, 40% passed the hand .

Table (6): Demonstrating the association between the Hand-cover test pass and Academic year

		Year				Total
		2014	2015	2016	2017	
Hand-cover test	**yes**	13	15	10	8	46
		65.0%	75.0%	50.0%	40.0%	57.5%
	no	7	5	10	12	34
		35.0%	25.0%	50.0%	60.0%	42.5%
Pearson Chi-Square value [5.934a], 2-sided P-value [0.115]						

Table 7 shows that the grammatical mistakes decrease between 2014-2017. In 2014, grammatical mistakes were present by 50%, in 2015 it increased to 65% then decreased in 2016 to 30% and in 2017 it became 25%. The Chi-Square Test value were found to be (8.389) with insignificant P-value (0.039).

Table (7): Demonstrating the association between the grammatical mistakes and Academic year

		Year				Total
		2014	2015	2016	2017	
Grammatical Mistakes	**yes**	10	13	6	5	34
		50.0%	65.0%	30.0%	25.0%	42.5%
	no	10	7	14	15	46
		50.0%	35.0%	70.0%	75.0%	57.5%
Pearson Chi-Square value [8.389a], 2-sided P-value [0.039]						

Table 8 shows there were no very long correct answer throughout the academic years.

Table (8): Demonstrating the distribution of the Very long correct answer through Academic years

		Year				Total
		2014	2015	2016	2017	
Very long correct answer	**no**	20	20	20	20	80
		100.0%	100.0%	100.0%	100.0%	100.0%
	yes	0	0	0	0	0
		0%	0%	0%	0%	0%

Cover test. The Chi-Square Test value were found to be (5.934) with insignificant P-value (0.115).

Table 9 shows that the repeated word in the stem and options decreased from 15% after 2014 to 5.0% and remained steady throughout the subsequent academic years. The Chi-Square Test value were found to be (2.162a) with insignificant P-value (0.539).

Table (9): Demonstrating the association between the repetition of words and Academic year

		Year				Total
		2014	2015	2016	2017	
Repeated word in the stem and options	**yes**	3	1	1	1	6
		15.0%	5.0%	5.0%	5.0%	7.5%
	no	17	19	19	19	74
		85.0%	95.0%	95.0%	95.0%	92.5%
Pearson Chi-Square value [2.162], 2-sided P-value [0.539]						

Table 10 demonstrated that merging more than one item in one answer increased from 10% to 15% in 2015 and remained steady, at 5% throughout the 2016 and 2017 academic years. The Chi-Square Test value were found to be (1.722a) with insignificant P-value (0.632).

Table (10): Demonstrating the association between merging more than one item in one answer in academic years

		Year				Total
		2014	**2015**	**2016**	**2017**	
Merging more than one item in one answer	**Yes**	2	3	1	1	7
		10.0%	15.0%	5.0%	5.0%	8.8%
	No	18	17	19	19	73
		90.0%	85.0%	95.0%	95.0%	91.3%
Pearson Chi-Square value [1.722a], 2-sided P-value [0.632]						

Table 11 demonstrates that long and exhausting options were only 5% during the 2016 academic year.

Table (11): Demonstrating the association between Long and exhausting options in Academic year

		Year				Total
		2014	**2015**	**2016**	**2017**	
Long and exhausting options	**yes**	0	0	1	0	1
		0.0%	0.0%	5.0%	0.0%	1.3%
	no	20	20	19	20	79
		100.0%	100.0%	95.0%	100.0%	98.8%
Pearson Chi-Square value [3.038a], 2-sided P-value [0.386]						

Table 12 illustrates that absolute terms were used very little, only 5% in 2014 and 10% in the 2017 academic year. The Chi-Square Test value were found to be (3.810a) with insignificant P-value (0.283).

Table (12): Demonstrating the association between using absolute terms and Academic year

		Year				Total
		2014	**2015**	**2016**	**2017**	
Using absolute terms	**yes**	1	0	0	2	3
		5.0%	0.0%	0.0%	10.0%	3.8%
	no	19	20	20	18	77
		95.0%	100.0%	100.0%	90.0%	96.3%
Pearson Chi-Square value [3.810a], 2-sided P-value [0.283]						

Table 13 shows that non-logical options were used more so in 2014 at 20% than decreased from 5% in 2015 and 2016 to 0% in 2017. The Chi-Square Test value were found to be (6.486a) with insignificant P-value (0.090).

Table (13): Demonstrating the association between using non-logical option and Academic year

		Year				Total
		2014	**2015**	**2016**	**2017**	
Using non-logical option	**yes**	4	1	1	0	6
		20.0%	5.0%	5.0%	0.0%	7.5%
	no	16	19	19	20	74
		80.0%	95.0%	95.0%	100.0%	92.5%
Pearson Chi-Square value [6.486a], 2-sided P-value [0.090]						

The Chi-Square Test value were found to be (3.038a) with insignificant P-value (0.386).

Table 14 demonstrates that poorly arranged numeric data was apparent in the 2015 and 2017 academic years at 5%. The Chi-Square Test value were found to be (2.051a) with insignificant P-value (0.562).

Table (14): Demonstrating the association between poorly arranged numeric data and Academic year

		Year				Total
		2014	**2015**	**2016**	**2017**	
Poorly arranged numeric data	**yes**	0	1	0	1	2
		0.0%	5.0%	0.0%	5.0%	2.5%
	no	20	19	20	19	78
		100.0%	95.0%	100.0%	95.0%	97.5%

Pearson Chi-Square value [2.051], 2-sided P-value [0.562]

Table 15 shows that negatively constructed items were 15% in 2014, 10% in 2015 then significantly increased after 2015 to 50%. The Chi-Square Test value were found to be (10.923a) with insignificant P-value (0.012).

Table (15): Demonstrating the association between negatively constructed items and Academic year

		Year				Total
		2014	**2015**	**2016**	**2017**	
Negatively constructed items	yes	3	2	8	10	23
		15.0%	10.0%	40.0%	50.0%	28.8%
	no	17	18	12	10	57
		85.0%	90.0%	60.0%	50.0%	71.3%

Pearson Chi-Square value [10.923a], 2-sided P-value [0. 012]

Table 16 demonstrates that phrases such as "None of the above "or "all of the above were most commonly used in 2015 at 25%. The Chi-Square Test value were found to be (9.237a) with insignificant P-value (0.026).

Table (16): Demonstrating the association between Using phrases like: "None of the above" or "all of the above" and Academic year

		Year				Total
		2014	**2015**	**2016**	**2017**	
Using phrases like: "None of the above" or "all of the above"	**yes**	1	5	0	1	7
		5.0%	25.0%	0.0%	5.0%	8.8%
	no	19	15	20	19	73
		95.0%	75.0%	100.0%	95.0%	91.3%

Pearson Chi-Square value [9.237a], 2-sided P-value [0.026]

Table 17 demonstrates that unclear or vague lead-in options were apparent throughout the academic years at 10% in 2015 and 2016, and at 15% in 2014 and 2017. The Chi-Square Test value were found to be (0.457a) with insignificant P-value (0.928).

Table (17): Demonstrating the association between "unclear or vague lead-in options "and Academic year

		Year				Total
		2014	**2015**	**2016**	**2017**	
Unclear or vague lead-in options	yes	3	2	2	3	10
		15.0%	10.0%	10.0%	15.0%	12.5%
	no	17	18	18	17	70
		85.0%	90.0%	90.0%	85.%	87%

Pearson Chi-Square value [.457], 2-sided P-value [0.928]

4.2 Discussion

MCQs have commonly served as summative assessment of choice in undergraduate medical education due to their convenient standardization, broad sampling of knowledge and and ease of assessing large groups of students. This study, which analyses written assessments for their cognitive level and presence of items writing flaws, coinsides with a recent study [1]. The present research found that 56.3% MCQs assessed the recall of isolated facts while the remaining 43.7% MCQs evaluated competence in data interpretation. There were no MCQ assessing application and analysis i.e. the higher cognitive domains. This may be explained by the fact that MCQs at recall level are easier to construct as they need less knowledge and less time investment as compared to problem solving MCQs, which require training and expertise [13].

Tarrant and Ware found when analysing a nursing examination that over 90% of MCQs addressed low cognitive levels, and that MCQs written at a lower cognitive level were far more likely to contain item writing flaws [15][19]. Jozefowicz et al. evaluated the standard of internally produced examinations at three US medical schools and noted that the overall quality of the questions was low. Several studies have confirmed that MCQs not only test the knowledge of the examinees but can also be used for measuring higher cognitive skills [10][20]. One of the main problems affecting the quality of MCQs is the presence of item writing flaws. Item-writing flaws (IWFs) result from nonobservance of accepted item-writing guidelines and can affect test-takers performance, making the items either more difficult or easier to answer [10] [2]. The present study found 132 IWFs over total of 80 MCQs. Downing assessed the quality of four medical examinations carried out in the United States of America, and found that 46% of MCQs contained IWFs. Downing noted that 10–15% of students who were classified as failures would have been classified as pass if items with IWFs were removed [11].

Results of the current study showed presence of flawed items could possibly be attributed to insignificant faculty development programs. Flawed items affect difficulty and discrimination index, low difficulty and poor discrimination in an item promotes low achievers. Higher difficulty can be achieved by reducing IWFs and improving cognitive levels of the test items. Another common factor that influences the validity and effectiveness of the MCQs test is the grammatical errors, which have a significant impact on the way the questions can be interpreted [4]. This research also proved that grammatical errors made up a total of 42.5% of MCQs. Since the matter of grammatical errors is mostly a language-based issue, language modification and revision are recommended, whenever possible, by language experts before the submission of tests. Moreover, negatively constructed items and hand cover tests constituted a high percentage among the other tested flaws of 28.8% and 42.5%, respectively.

These results thus illustrate that there is deficient knowledge and skill about how well-built and valid MCQs are constructed. A possible justification of such issues could be that medical education is still a newly enlisted program and needs further enhancement among university staff members. Many examiners refer back to past examination paper questions using Q-banks without verifying the accuracy and validity of the items [15].

The present study suggests that there is a need to improve the quality of our assessment tools because if the assessment tools measure low cognitive level, it will not only decrease the validity of the exam but also encourages the students to settle on surface learning.

Downing suggested the use of a test blue print [11]. A blueprint is a crucial measure in producing a valid and reliable test. It can be as simple as a chart or a table that lists the objectives of the course and the weighting of each component. A blueprint helps the exam writer to allocate an accurate percentage of questions to each content area at a desired cognitive level. Tarrant highlighted that removing IWFs from MCQs does not necessarily change the cognitive domain of a question, but writing questions at higher cognitive levels inherently removes numerous IWFs [19].

Limitations of the study: The study analyzed results of only one module, and students' scores in only one subject. Moreover, difficulty and discrimination indices were not available

V. CONCLUSION

The cognitive level of MCQs as an assessment tool is low and IWFs are very common. Further on, effective item construction requires competence and awareness of item-writing principles. Assessment is a crucial part of the learning process and educators should keep in mind it is one of the main factors that influences the students' approach to learning and their future learning goals. Therefore, due care and attention should be afforded to training item writers and creating valid and reliable test items.

RECOMMENDATIONS

The present study suggests that it is imperative to improve the quality of our assessment tools. Assessment tools designed to measure low cognitive levels will not only decrease the validity of an exam but also compel the students to adopt surface learning approaches which are ultimately neither desirable nor sustainable. The present study is done on a single module, it is recommended to cover all subjects and modules to achieve a more reliable result.

REFERENCES

[1]. Access, O., 2017. Evaluation of cognitive levels and item writing flaws in medical pharmacology internal assessment examinations. , 33(4), pp.866–870.

[2]. Access, O., 2014. Evaluation of Multiple Choice and Short Essay Question items in Basic Medical Sciences. , 30(1), pp.3–6.

[3]. Access, O., 2013. Identification of technical item flaws leads to of the quality of single best Multiple Choice Questions. , 29(3), pp.715–718.

[4]. Al-shaikh, G.K. et al., 2015. Faculty development programs improve the quality of Multiple Choice Questions items ' writing.

[5]. Ayoob, A.R. & Williams, L.E., 2015. Prevalence of Flawed Multiple- Choice Questions in Continuing Medical Education Activities of Major Radiology Journals. , (April), pp.698–702.

[6]. Barzegar, M. et al., 2014. Comparison of Multiple-Choice Questions in Quality Parameters of Pediatric Residency Tests between the Pre-Board Examination of Tabriz University of Medical Sciences and National

Board Examination in 2007 and 2011. , 3(1), pp.31–37.

[7]. Case, S.M. & Swanson, D.B., Constructing Written Test Questions For the Basic and Clinical Sciences Third Edition.

[8]. Coughlin, P.A., 2017. How to Write a High Quality Multiple Choice Question (MCQ): A Guide for Clinicians. European Journal of Vascular & Endovascular Surgery, pp.1–5. Available at: http://dx.doi.org/10.1016/j.ejvs.2017.07.012.

[9]. Developing, I. & Items, M., 2005. Technical Guidelines for the Construction of Multiple-Choice Questions Including.

[10]. Dibattista, D. & Kurzawa, L., 2011. Examination of the Quality of Multiple-choice Items on Classroom Tests Examination of the Quality of Multiple-choice Items on Classroom Tests. , 2(2).

[11]. Downing, S.M., 2005. The Effects of Violating Standard Item Writing Principles on Tests and Students : The Consequences of Using Flawed Test Items on Achievement Examinations in Medical Education. , pp.133–143.

[12]. Education, E. et al., Universititunkuabdulrahman,

[13]. Freiwald, T. et al., 2014. Pattern recognition as a concept for multiple- choice questions in a national licensing exam. , pp.1–6.

[14]. Item, T., 2013. Multiple-Choice Item Review Checklist. , pp.17–18.

[15]. Karkal, Y.R. & Kundapur, G.S., 2016. Item analysis of multiple choice questions of undergraduate pharmacology examinations in an International Medical School in India. , 5(3), pp.183–186.

[16]. Odriguez, M.C. & Group, F., 2017. Book Review. , 2(1), pp.101–104.

[17]. Rush, B., 2012. Examination Item Development Multiple - choice Questions Stem errors. , pp.1–15.

[18]. Zimmaro, D.M. & Ph, D., 2004. Writing Good Multiple-Choice Exams. , (512).

[19]. Tarrant M, Ware J., 2012. A framework for improving the quality of multiple-choice assessments. Nurse Educ. May-Jun;37(3):98-104. doi: 10.1097/NNE.0b013e31825041d0. PMID: 22513766.

[20]. Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., and Glew, R. H., 2002. The quality of in-house medical school examinations. Acad. Med. 77, 156–161. doi:10.1097/00001888-200202000-00016

University of Gezira Faculty of Medicine Education development and Research center Master Program in Health Professional Education. A validated check list based on the guidelines of the NBME (National Board of Medical Examinations) for evaluate the Quality of Multiple Choice Questions in Fixed Prosthodontic module, Faculty of Dentistry, Gezira University (2014-2017)

Exam No			
Assessed Cognitive Domain	**1-** **Knowledge** (recall)		
	2- **Comprehension**		
	3- **Application**		
	4- **Analysis**		
	5- **Synthesis**		
	6- **Evaluation**	Yes	No
Lead-in	In contextual MCQs: the question can be answered even if the options are covered “*hand-cover test: pass*”		
Options should be free from:	**General flaws:**		
	Grammatical Mistakes.		
	Very long correct answer.		
	Repeated word in the stem and options.		
	Merging more than one item in one answer.		
	Long and exhausting options.		
	Using absolute terms.		
	Using non-logical option.		
	Flaws contribute to irrelevant difficulty:		
	Poorly arranged numeric data.		
	Negatively constructed items.		
	Using phrases like: “None of the above” or “all of the above”.		
	Unclear or vague lead-in or options.		