Simplified Guidelines for Multiple-Choice Question Writing to Increase Faculty Compliance and Ensure Valid Student Results

Magdy Hasan Balaha

Faculty of Medicine, Tanta University, EGYPT

ABSTRACT

This study aimed at presenting a simple, compiled multiple-choice questions' (MCQs) guideline to the busy faculty to increase their compliance, to improve the item-writing quality and to enhance valid assessment. The current published MCQs guidelines were examined, preferably those in the field of medical education, from different medical schools in the USA, Canada, Britain, Europe, Australia and the Arabic area. Searching databases and publications were done through the Egyptian Knowledge Bank. Some of the guidelines were downloaded from ResearchGate or Google Scholar. After applying selection and exclusion criteria, 29 documents were legible and lastly only 14 guidelines were included in the final review construction. The data was cross-mapped to evaluate the shared points. Similar points were added together. A common single frame was made from which a simplified shortlist was prepared. The list included 25 criteria that were assigned into four areas such as the item format, item content, stem construction and alternative writing. Adding or re-allocation of some points was made to reach to the compiled form. The compilation and simplification were done to synthesise a 20-point list; five in each section. This list was presented in a table form and as a designed coloured card. The simplified shortlist is a single-page guide and the coloured pocket card is a novel product. They compile the scientific content of the guidelines; moreover, present them in an easier and simpler way to the busy faculty. Their adoption might be a good asset for faculty compliance, improving item-writing quality, and enhancing the assessment process; aiming at ensuring valid student results.

Keywords: *Multiple-choice questions, Item-writing flaws, Item-writing guidelines, Test validity*

CORRESPONDING AUTHOR Professor Magdy Hasan Balaha, Faculty of Medicine, Tanta University, Al Geish St. Medical Campus, 31527, Egypt | Email: magdy.balaha@med.tanta.edu.eg

INTRODUCTION

Assessment provides valid information, concerning the progress and attainment of the expected learning outcomes. Since the assessment is the main drive of student learning, the methods of assessment determine the students' roadmap towards learning. The students tend to utilise a superficial approach when assessment emphasise on simple recall of knowledge; while they adopt a deeper approach if the assessment demands higher levels of cognitive abilities (1, 2).

Multiple methods are used to assess medical students. Testing methods are planned in order not to confuse the students; moreover, to yield accurate scores that reflect the actual level of knowledge acquisition. Multiple-choice questions (MCQs) are appropriate for measuring knowledge, comprehension and could be

designed to measure higher cognition such as application and analysis. Even those MCQs produced by experienced item writers, it may still have some item flaws. Faculty training in test item construction is crucial to make high quality exam items. Exam vetting process would ensure a high standard of test items and sustain their content validity (3). Downing reported that MCQs item flaws had a great drawback on the item validity (4). Well-constructed high quality MCQs could allow for the evaluation of a wide range of scientific content and the avoidance of technical items flaws. Both are essential to improve reliability, validity and ease the scoring interpretation (5–6).

Many guidelines were published. The primary taxonomies were published by Haladyna and Downing; and Haladyna, Downing and Rodriguez (7–8). The National Board of Medical Examiners (NBME) issued the most detailed source of guidelines for medical faculty. They issued the first edition in 1996 and the fourth edition in 2016 (9–10). With the plethora of guidelines, many training programmes were vetted at each university. However, virtually all the publications discussing the analysis of MCQs' performance had reported a high percentage of item flaws and non-functioning distractors (11–12). Meaningful interpretation of assessment scores requires content-related validity, construct validity and the item quality parameters. The assessment flaws or misinterpretation would endanger the truth about actual student learning. Some authors had defined the validity as the evidence generated from various sources to support the meaning assigned to scores, obtained on an assessment instrument or a test (13–14).

Writing a quality single best answer (SBA) MCQs needs training and experience with timed feedback to the teaching faculty. Increased number of non-functional distractors indicates the difficulty in developing the alternatives by the teaching faculty. This would be reflected in poor test validity and reliability (15).

There are visible paradoxes regarding MCQs such as many MCQs guidelines, high percentage of item flaws and increased non-functioning distractors. Many factors might be inflicted in these paradoxes. Of the speculated reasons: the detailed description of theoretical guidelines, the educational terminology, and the debates and arguments. This might overwhelm the busy faculty. These factors may lead to testing bias and hence, questionable assessment scores. Moreover, the administrative leadership and approach to this aspect need to be more decisive. Another possible source for that bias could be initiated by using others' MCQs or items; whether supplied by the commercial question banks, distributed with textbooks, or presented with online versions of some books (16–17).

Medical faculty in the colleges of medicine, dentistry, pharmacy, health sciences and nursing are very busy in their practice and patient services. They are faithfully good teachers. Meanwhile, the assessment that is composed of a scientific part as well as an art needs a lot of concern and compliance to guidelines. The science part is controlled and well-managed by the faculty; however, the art of assessment design and construction still needs compliance to the guidelines, proper application and continuous rehearsals. Non-compliance is risky to the exam validity and reliability; as well as a possible questionable knowledge and abilities of the medical graduate.

The purpose of this study was to present a simple and compiled MCQs guideline to the busy faculty to increase their compliance, to improve the item-writing quality and to enhance valid assessment.

METHODS

This study was done by reviewing the literature for the primary guidelines in item writing, then stressing on the medical examiner's guidelines, and finally making a comparison list, to sort the shared points, with their supporting evidence. After

that, studying of the findings were done from the perspective of medical concern, simplicity and applicability. This study is a comprehensive review of the available or published guidelines. It is not a systematic review.

Search for Sources of the Guidelines

1. The studies that discussed item construction guidelines were considered eligible. Assessment of these studies were done before inclusion in this review.

2. The inclusion criteria of the articles and guides were: (a) guidelines discussing most of the item components, (b) widely cited by other selective articles, (c) being issued or endorsed by medical organisation, and (d) representing different international medical education schools.

3. The search for relevant studies was performed using the following databases: MEDLINE, PubMed and ERIC. In addition, EBSCOhost, Elsevier, Sage, Springer, Taylor & Francis, Wiley, and JSTOR were searched (18). Due to the limited numbers of articles, some of the guidelines were downloaded by searching the ResearchGate and Google Scholar as well as from the medical schools and organisations' websites.

4. The used keywords in the search included MCQs, item-writing flaws, item-writing guidelines, test validity, review and meta-analysis. Using AND, OR; "USA, Canada, Britain, Europe, Australia and the Arabic area" and "medical assessment guidelines".

Screening and Revision of the Findings

1. A total of 29 retrieved studies as guidelines were examined. The articles fulfilling the inclusion criteria were included. The excluded articles were either presenting a very brief rules without the explanatory background, just summarising another guide or based on small studies.

2. A total of 14 studies met the selection criteria and were included in this review in Table 1 (7–10; 19–28).

Master File of Data Comparison

1. The primary filters were the two MCQs taxonomy and guidelines; published by Haladyna and Downing (7), and Haladyna, Downing and Rodriguez (8). These reports divided MCQs writing into six areas: general procedural item writing, content concerns, stem construction, general option development, correct option development and distractor development.

2. An excel spreadsheet was constructed for making a comparison scheme, based on the six main areas and their sub items, as proposed by Haladyna and Downing (7).

3. All subsequent MCQs guidelines, taxonomies or technical reports were transferred to the master file.

Table 1: The final included and analysed guidelines for the review

G1	G2	G3	G4	G5	G6	G7
Haladyna 1989 (7)	Haladyna 2002 (8)	Case 2002 (9)	Collins 2006 (19)	Tarrant 2006 (20)	Wood 2004 (21)	SCFHS 2011 (22)
G8	**G9**	**G10**	**G11**	**G12**	**G13**	**G14**
Brame 2013 (23)	Schuwirth 2014 (24)	Tenore 2015 (25)	Davis 2016 (26)	Zimmaro 2016 (27)	CLIME 2017 (28)	Paniagua 2016 (10)

Cross-mapping

1. The secondary filter, in the comparison of all guides was the NBME guidelines (9). Any point in other guidelines were validated versus this guide. Some other crucial points were also included.

2. Similar points in different sub-headings were collectively unified. Some points with very academic and difficult meaning were replaced by simple words.

3. A consensus of the statements that must be followed; and the points that must be avoided in MCQs writing were reached.

4. Final re-arrangement of the comparison list was made to make a logical sequence of steps in MCQs writing.

Tabulation

1. Final tabulated forms in different areas of MCQs writing were made from the comparison list and supported with evidence; arranged from G1 to G14 (7–10; 19–28).

2. The frequency (percentage) of mentioning of each point in all studied guidelines was added.

3. These tables are:

   a. Table 2 showed item format guidelines and item-content guidelines. For the item format guidelines, the most frequent were the choice of SBA and item independence; followed by English writing (simplicity, grammar and language), and lastly, the vertical format. As regards item-content guidelines, the most frequent were linking each item to important topics or concepts, or learning objectives, items without tricks, and then the item's ability to measure level two of thinking.

   b. Table 3 showed stem construction guidelines. The highest percentage were inclusion of most of the content in the stem, the style of vignette and positivity in wording.

   c. Table 4 showed alternative writing guidelines. This part had many points which represent sound distractors (plausible and homogenous), difficulty item flaws (absolute or vague terms, combinations, none of the above) and test-wiseness flaws (repetition, grammar and convergence).

Synthesis of a Shortlist

1. A shortlist of all components in the previous three tables were looked at as per the highest rank in their percentage of frequency.

2. The statements of the 20 items were simplified and reduced to be suitable for designing the card.

3. Table 5 showed synthesised shortlist of guidelines. A collective 20-statements were formulated; five in each section. Some points were mobilised or joined.

Pocket Card

1. Double-faced pocket card was formatted with the help of a designer. The folded card will be in the size of some mobile phones, to be suitable to be put in the pocket of the assessor or examiner, during time of test construction.

2. Card dimensions: It will be 17 cm length and 17 cm width. After folding, it will 17 cm by 8.5 cm. Anterior or front face will contain the scientific information. The front surface will be folded. The back or exterior of the card, upon folding will be two sides. Each is 8.5 cm.

3. The interior part will contain the statements of the simplified rules. The exterior or back will contain the references on one side, and the publication and general information on the other side.

Table 2: Distribution of item-format and item-content guidelines as extracted from 14 of the large publications or guidelines

#	Rule	Item-writing (format)	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	%*
1	Use	Single best answer.	✓	✓	✓		✓	✓	✓		✓	✓		✓	✓	✓	71
2	Use	Format the item vertically.	✓	✓	✓		✓		✓					✓		✓	36
3	Use	Sound English grammar and spelling.	✓	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓		64
4	Use	Clear simple understandable vocabulary.	✓		✓		✓		✓	✓		✓	✓	✓	✓	✓	64
5	Avoid	Cueing to or hinging with other items.	✓	✓		✓			✓	✓	✓	✓	✓	✓	✓	✓	71
6	Avoid	The complex multiple-choice.	✓		✓		✓	✓	✓	✓				✓	✓	✓	57
7	Use	Test for important or significant material.	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	93
8	Use	Base each item on a single objective.	✓			✓	✓	✓	✓			✓	✓	✓	✓	✓	71
9	Use	Own novel material and consider culture, gender issues.	✓	✓							✓			✓	✓		36
10	Use	Measure different levels of thinking.	✓		✓	✓						✓	✓		✓	✓	50
11	Avoid	Tricky or misleading items.	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓	✓	71

Note: %* = Percentage of the presence of this item in all the guidelines from G1–G14.

Table 3: Distribution of stem construction guidelines as extracted from 14 of the large publications or guidelines

#	Rule	Stem construction	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	%*
12	Use	State the stem as a question or a vignette followed by lead in as question. If completion form is necessary, only at end of stem.	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	86
13	Use	Word the stem (or lead in) positively without negative terms; avoid negative phrasing, unless highly indicated with specific rules.	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	86
14	Use	Include the central idea and most of the phrasing in the stem.	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	86
15	Use	Cover-the-options rule to answer the question without looking at options.			✓			✓	✓				✓	✓	✓	✓	50
16	Avoid	Absolute terms (never, always, usually, often, only, all).	✓		✓	✓	✓	✓	✓			✓	✓		✓	✓	64

Note: %* = Percentage of the presence of this item in all the guidelines from G1–G14.

Table 4: Distribution of alternatives or options guidelines as extracted from 14 of the large publications or guidelines

#	Rule	Alternatives or options	G1	G2	G3	G4	G5	G6	G7	R8	G9	G10	G11	G12	G13	G14	%*
17	Use	Plausible & logical distractors.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100
18	Use	Keep all options homogeneous in content or domain.	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	86
19	Use	Place options in logical, chronological, or numerical order.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100
20	Use	Keep options independent; without overlapping.	✓	✓	✓	✓		✓	✓	✓	✓				✓		64
21	Use	Keep the length of options consistent.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100
22	Avoid	Using "all of the above" or "A & B."	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	93
23	Avoid	Using "none of the above."	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	93
24	Avoid	Vague terms (frequently, occasionally, rarely, commonly, can, may).	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓	79
25	Avoid	Test-wise item flaws; as faulty grammar, repetition and convergence.	✓	✓	✓		✓		✓	✓	✓		✓	✓	✓	✓	79

Note: %* = Percentage of the presence of this item in all the guidelines from G1–G14.

Table 5: Synthesised list of guidelines as extracted from 14 of the large publications or guidelines (14 points to be done and 6 points to be avoided)

#	Rule	Item format	%*
1	Use	Single best answer	71
2	Use	Format vertically	36
3	Use	Sound English	64
4	Use	Clear simple vocabulary	64
5	Avoid	Cueing or hinging	71
		Item content	
6	Use	Important significant material	93
7	Use	Single objective per item	71
8	Use	Own novel material	36
9	Use	Different thinking levels	50
10	Avoid	Tricky items	71
		Stem construction	
11	Use	Stem with vignette and question	86
12	Use	Positive stem and lead-in	86
13	Use	Central idea in the stem	86
14	Use	Cover-the-options rule	50
15	Avoid	Absolute terms	64
		Alternatives or options	
16	Use	Plausible, homogeneous options with parallel length	95
17	Use	Options in logical order without overlapping	82
18	Avoid	AOTA, NOTA or "Complex form"	81
19	Avoid	Vague terms	79
20	Avoid	Test-wise item flaws	79

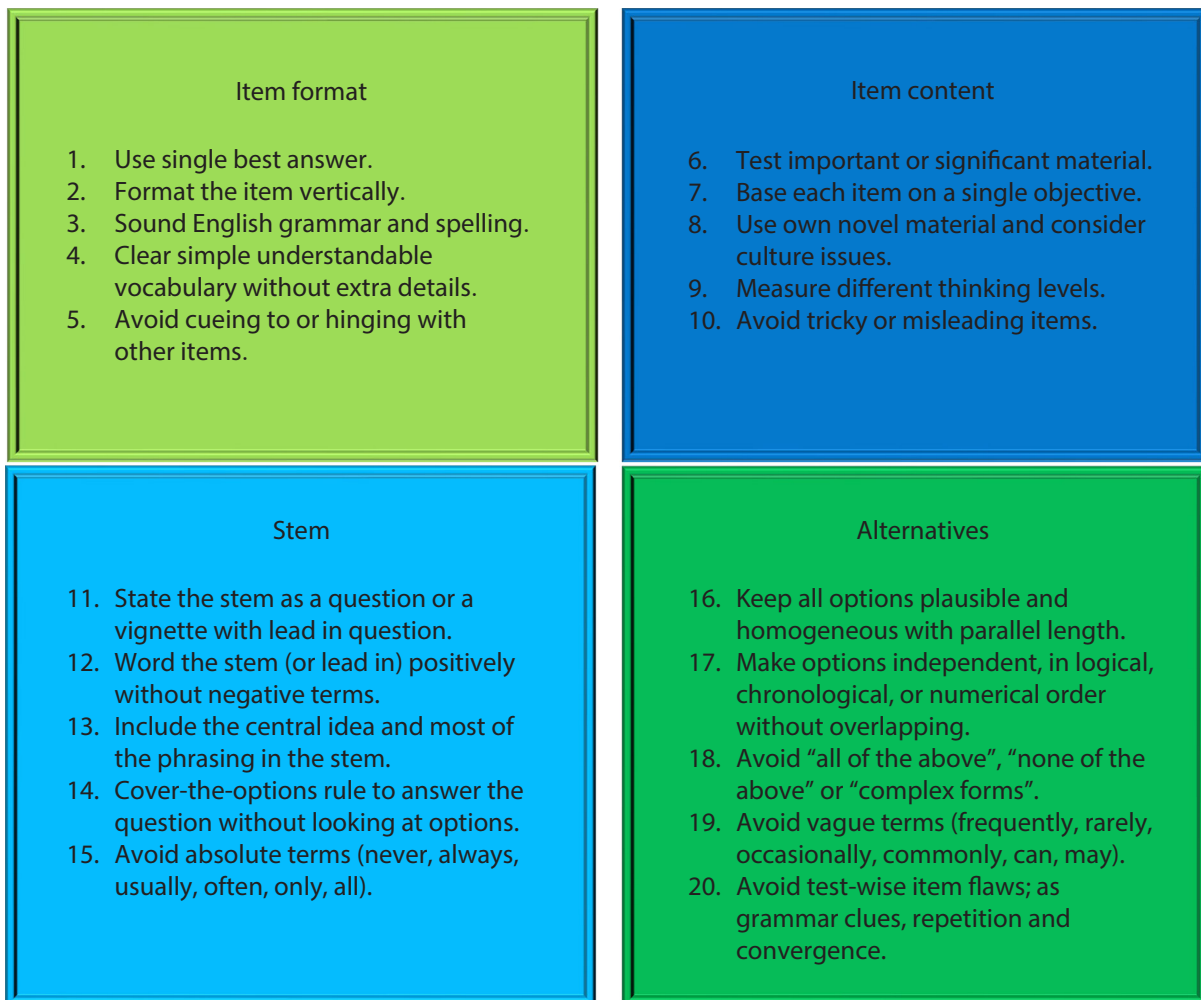Note: %* = Percentage of the presence of this item in all the guidelines from G1–G14.

4. Figures 1 and 2 showed the artistic design of the card, front face and cover face. Figure 1 contained the front face with all boxed guidelines and Figure 2 has a cover contained the references and publication.

RESULTS AND DISCUSSION

In spite of the available MCQs guidelines, the reported prevalence of poor item quality in most of the publications is high. Biased or flawed items may lead to either inflated or underestimated results. This is dangerous for exam validity and reliability. This was ascribed partially to decreased faculty compliance to the available guidelines. Thinking of a more suitable and attractive method was the motive for this novel work. The purpose was not to make a new guideline. It was aiming to compile a simplified scheme to be published in a pamphlet like coloured card to increase the compliance.

The main four headings of item writing are item format, item content, stem and alternatives will be discussed.

Item format	Item content
1. Use single best answer. 2. Format the item vertically. 3. Sound English grammar and spelling. 4. Clear simple understandable vocabulary without extra details. 5. Avoid cueing to or hinging with other items.	6. Test important or significant material. 7. Base each item on a single objective. 8. Use own novel material and consider culture issues. 9. Measure different thinking levels. 10. Avoid tricky or misleading items.
Stem	Alternatives
11. State the stem as a question or a vignette with lead in question. 12. Word the stem (or lead in) positively without negative terms. 13. Include the central idea and most of the phrasing in the stem. 14. Cover-the-options rule to answer the question without looking at options. 15. Avoid absolute terms (never, always, usually, often, only, all).	16. Keep all options plausible and homogeneous with parallel length. 17. Make options independent, in logical, chronological, or numerical order without overlapping. 18. Avoid "all of the above", "none of the above" or "complex forms". 19. Avoid vague terms (frequently, rarely, occasionally, commonly, can, may). 20. Avoid test-wise item flaws; as grammar clues, repetition and convergence.

Figure 1: The artistic design of front face of the pocket card with all boxed guidelines (17 cm × 17 cm, after folding; 17 cm × 8.5 cm).

Item Format

Single best answer (SBA) (71%)

There are many types of MCQs. SBA question is more logic in medical practice, if compared to the selection of the correct answer because there is some degree of correctness in most of the medical situations. Justifying the, "best choice", or the keyed option may be challenging to the student as it stimulates higher thinking domains. SBA is supported by many guidelines. NBME stated clearly that all exams under their supervision would be all SBA (7–10, 20–22, 25, 27, 28).

Vertical format (36%)

There are three types of MCQs designing which are horizontal, square and vertical. Vertical arrangement is better as the student can skim the answers easily and compare the arranged or ordered choices to select the best one (7–8, 16, 27). In the horizontal lists, students need to span a larger area, therefore, the user must move their focus larger distances, which is tiring to the eyes. The same happens with the square style where the user must apply a Z-pattern to scan the list, moving from left to right and back to left (see the examples).

What is the best likely diagnosis?
a. Brucellosis
b. Hepatitis
c. Malaria
d. Typhoid

What is the best likely diagnosis?
a. Brucellosis
b. Hepatitis
c. Malaria
d. Typhoid

What is the best likely diagnosis?
e. Brucellosis
f. Hepatitis
g. Malaria
h. Typhoid

References

1. Haladyna & Downing. (1989).
2. Haladyna, Downing & Rodriguez. (2002).
3. Case & Swanson. NBME guidelines. Revised 3rd ed. (2003).
4. Paniagua & Swygert. NBME guidelines. 4th ed. (2016).
5. Collins J. Writing multiple-choice questions. (2006).
6. Tarrant M. The frequency of item writing flaws. (2006).
7. Wood. Developing MCQs for RCPSC Certification Examinations. (2006).
8. SCFHS item writing manual. (2011).
9. Brame C. Writing good MCQs. vanderbilt.edu. (2013).
10. Schuwirth L. Australian quality of assessment items. (2014).
11. Tenore A. Guide to successfully writing MCQs: UEMS-CESMA guideline. (2015).
12. Davis M. Joint Committee on Intercollegiate Examinations: Examiners' Manual. (2016).
13. Zimmaro DM. Writing good MCQs. Texas, Austin. (2016).
14. CLIME. Writing MCQs. Washington School of Medicine. (2017).

MCQ Pocket Card
"MPC"
Compiled Guidelines
by
Magdy Balaha
Tanta University
Egypt
2019
Published in
EIMJ Vol.11, No.4, 2019: Q6

Figure 2: The cover page sample of the pocket card contained the references and publication (17 cm × 17 cm, after folding; 17 cm × 8.5 cm).

Sound English (64%)

This point is very crucial. Each word or word combinations may have a different meaning if punctuation marks were not used accurately. The context of the question, "past, present or future", needs to be considered. If the stem was in the past and the question is in the future tense, this may deviate the thinking of the good student to another context. In addition, the grammar and spelling of the stem must be consistent with the options to avoid any reference to the answer or clueing (7–10, 20, 22, 25, 27).

Clear simple vocabulary (64%)

There is no need for unnecessary information as this will abuse the time and distract thinking. Difficult wording is not preferred in formulation of questions. Language barriers for nonnative English speakers must not affect the student abilities in the exam. Poorly worded or ambiguous questions can confuse even the knowledgeable students and cause them to answer incorrectly. If the faculty members applied some of the pre-validation steps as peer reviewing and pilot testing, lot of these points can be highlighted and avoided (7–10, 20, 22, 25–29).

Avoid clueing, cross-referring or hinging (71%)

Clueing is the presence of relational point in the syntax or the grammar, which may reveal an inference to the answer. If the answer will be inferred from another question, it is cross-reference. Hinging is the linking of one item's answer by answering of a previous item. The question stems must be revised to avoid answering of many other questions. All these flaws can be controlled, if the blueprint of the exam was well formulated in detail, up to the level of outcomes, its related content, level of the question; as well as the final test revision. These item flaws are unfair for the knowledgeable students, allow for testwiseness without studying and decrease the test validity (7–9, 20–22, 28).

Item Content

Important significant material (93%)

Any test is advisable to cover the breadth and depth of the knowledge to reach to the desired validity and relevance. It must sample the important points. Relevance is known by the course outcomes, which are important for the future practice. The students usually study what the teachers are inspecting; i.e. "what is coming in the exam". If we included all the major significant units, we are encouraging more deep learning and prepare the potential for future achievement. Doing that, we are increasing the different aspects of the exam validity: content, construct and predictive (7–10, 20–22, 24–28).

Single objective per item (71%)

Each question must have one domain that is decided by the action verb in the course objectives. Course specifications would enable the students to ascertain the nature and level of the content to be tested and would facilitate the faculty to create focused questions related to the outcomes. By doing this, we can achieve course outcomes, which are matched to the programme outcomes and increase the test predictive validity, reliability and impact (7, 9, 19, 22, 25–28).

Use your own novel material (29%)

Some questions in the test may not be related to a specific objective if they were imported from others' work as books or test banks. Alignment of these rented questions, with the course materials, and outcomes are crucial. Another point is the ethical issues regarding the copyrights of the questions' authors and legality of using without permission. Many drawbacks were observed in the rented questions such as the language level, cultural differences, style of the question and most importantly the improper quality and item flaws that are divulging MCQs guidelines (7–8, 20, 24–28).

Balance the different levels of thinking (50%)

Many classifications are present. One practical taxonomy is the reduced Bloom's taxonomy. The cognitive domain is collapsed it into three levels. Level I covered knowledge and recall of information, Level II covered comprehension and application, understanding and the ability to interpret data, and Level III tested problem-solving, the use of knowledge and understanding in new circumstances (20, 30).

MCQs format can allow the examiner to either test recalling of factual material, comprehension or to test higher-level of thinking as application, analysis and evaluation. MCQs should test at the same level of learning as the course objective. To simulate the future practice after graduation and increase the test impact and validity, it is recommended to keep specific percentages of questions, for both levels of thinking. One large Norwegian study stated that the MCQs exams must be at least 50% Level II (31) while, Washington School of Medicine, specified 80% of MCQs to be Level II (28). Rewording the questions rather than using the words and statements of the books or lectures, will also increase the level of the thinking. Moreover, the SBA question by itself, will challenge the discrimination ability and can test higher-order thinking. Recall type questions may begin by citing a disease and then asking what patient findings are expected. If we made the reverse by presenting a case scenario, followed by the question asking about the most likely diagnosis or management, we are formulating a higher-level question (7–10, 20, 25, 26, 28, 31).

Avoid tricky items (71%)

Testing in one of its aspects, aims at identifying the student academic level. If the exam proved that the student gained the needed knowledge, that is very sufficient and satisfactory. Misleading words or tricky complexity in the questions, can distract the student thinking and affect the test results. These are named as confounders. They are neither required nor having any educational benefit. Tricks, may also include the extreme difficulty, multiple negativity in the stem, vague or absolute terms, trivial or odd content, multiple correct answers, putting some irrelevant distractors or use complex logic questions known as K-type. All these practices must be avoided (7, 8, 10, 19, 22, 23, 25–28).

The MCQ Stem

Stem with a scenario and question (86%)

In recall type MCQs, we can use a very short question carrying small amount of information. Higher thinking questions may need a concise scenario (vignette), followed by the question (lead in). Short simplified scenario followed by an application or analytical question, is recommended to measure higher thinking skills; yet, some post graduate level MCQs may depend in longer scenarios. Whatever the type of vignette or scenario, the question must be relevant and answered only by reading the vignette. If the question can be answered without the scenario, this means an irrelevant case and must be corrected.

The items in the form of a question or a stem followed by question are preferable than partial sentence completion forms. They allow the student to focus on the answer, rather than holding the partial sentence in the working memory to find a completion, and hence causing his mental fatigue. Lot of guidelines omit the use of unfocused questions as "which of the following statements is correct?", "which of the following is true [or false]?". They are usually followed by heterogeneous statements as options. It is advised by the NBME to be replaced by "which of the following is the most likely finding?", "which of the following is the best likely treatment?", or "which of the following is the best appropriate investigation?". Many publications are against the completion form including NBME guidelines. If it was difficult for the faculty to formulate a question, the open or incomplete statement format could be used, provided that all other rules are respected (7–10, 20–23, 26–27, 32).

Positive stem and question (or lead-in) (86%)

Rodriguez (32) reported that the students often have difficulty-understanding items with negative phrasing as (not, except, unless or incorrect). Negative prefix such as uncharacteristic and unimportant are also similar. They poorly assess actual gained knowledge and require the students to shift their mental set from the positive to the negative; a task that may fail or being very tiring. In addition, the negative switching is harder for some people in nonnative English speakers (similar to "is it?" and "isn't it?"). Therefore, bad results on this type of questions might be due to the language rather than knowledge. This is one of the construct irrelevant errors.

If it was unavoidable situation to assess critical areas as mortality and disasters, the negative component must be in the stem. It is neither recommended in the question, lead-in, nor in the options. In these items, the options should be very short with absolute degree of correctness. Negative words should be emphasised with italics, bold face and capitalisation. The overall percentage of negative items in the exam should not represent more than 10% (7–10, 19–23, 25–28, 32).

Central idea in the stem (86%)

The stem should be meaningful by itself and present a definitive problem without any repetitive information that might appear in the options. We need to avoid the MCQs that is associated with unfocused stems, which do not make a clear context. The NBME defined some types of wording errors (verbosity errors) to be avoided by the item writers because they add another irrelevant difficulty to the exam. These included the extra words without benefit, the words which are unnecessary to answer the item and words carrying a misleading meaning (7–10, 19–22, 25–26, 28).

Cover-the-options rule (50%)

The rule is applicable for either straight questions or scenarios followed by questions. The students should be able to anticipate the answer of the question, without reading the options. By trying this, it is very helpful for the knowledgeable students without mental strain. It is advisable for item writers to test their items by covering up the responses while reading the question. However, sometimes we may apply the cover rule; however, the answer may not be found in the options. In this case, item writers are advised either to change the options to answer the question or to rewrite the stem or the question (9–10, 21–22, 24, 26, 28).

Avoid absolute terms (64%)

As stated previously in Item Content – avoid tricky items, the absolute and vague terms are not preferred in the stem or in the question. They include many words such as rarely, probably, never, ever, usually, and absolutely. Using words as likely, most likely, most effective and appropriate are preferred. There are certain wording style and a list of examples of the format for all types of questions; as reported in the NBME guidelines (7–8, 10, 19, 20–22, 25–26, 28).

Alternatives or Options

Plausible, homogeneous options with parallel length (95%)

All options, must be looking as reasonable possible answers (plausibility)

Developing reasonable (plausible) options (distractors) to the correct answer is of great importance for a high-quality test and to avoid test wiseness without studying. They are specifically formulated to be partially or apparently correct. So long we are using the SBA type, these reasonable options should attract the non-studying

students. If the options were totally implausible, the students will exclude them without any effort and the differential ability of the items will decrease. This give the less knowledgeable student more chances of guessing without studying the material. The exam quality will be lower in discrimination power as well as the decreased exam validity. Teachers who are simulating some of the test banks or whom are obliged to use a fixed number of four or five options in medical situations, often add some implausible options, that are known as "fillers". There are certain medical conditions, where the natural or logical scenario is only two options (e.g. types of movement in the interphalangeal joint; flexion and extension) or three options only (e.g. types of surgical interference; open, laparoscopic or robotic). In these conditions, any trial of bringing more options, will be using the fillers. Nowadays, many publications after Rodriguez (32) meta-analysis reported that the 3-option multiple choice items are very suitable without significant effect on item reliability; with very efficient distractors compared to four and five options.

Homogeneity

So long the item is adopting the question format or there is a question at the end of the stem; this needs a specific answer which mandates all the answers to be within the same scope or domain. Semantic homogeneity states that options must share common characteristics; all diagnosis, treatment or investigations, prognoses, or disposition. Grammatical homogeneity must be verified. All options are nouns, verbs, adjectives or sentences. The problem in completion form is that it can be followed by a group of heterogeneous options, which is not recommended. Non-homogenous options represent a major threat and considered as one of the testwiseness clues without studying. They are easily either excluded or selected, without studying, hence affect the item and exam validity (19–22).

Option length and structure

Some item writers often make the correct option longer and include more information to ensure that it is unambiguously correct. Others may include more detail in the correct answer to ensure that it is clearly the best choice. When students are unsure of the correct answer, a common wiseness practice is to select the longest or detailed option. This is called the "too long to be wrong" rule. Unfortunately, testwise students can use this to their advantage without studying. Knowledgeable students may be harmed, as they tend to read and analyse every option (7–10, 19–20, 22–28).

Arrange options without overlapping (82%)

To avoid any bias, alternatives must be arranged using certain character. Either alphabetical, chronological or logical is accepted. This arrangement will put the right answer at any position without any need to manually randomise the position of the key. If numbers are used, arrange them ascending or descending. In this regard, it is also recommended to avoid putting eccentric values or non-logical numbers as these are excluded from the start. On the other hand, options must be written in a way to be mutually exclusive. No overlap is permitted whether semantic (partial and total meaning) or true overlap between arithmetic parameters. In SBA type of MCQs, overlap divulge the rule of selecting the best. Alternatives with overlap are "tricky" items to the good students and affect the validity and trust for the testing process (7–10, 19–28, 32).

Avoid AOTA, NOTA or "complex form" (81%)

AOTA (All of the above)

AOTA is used on a non-justifiable basis. Students can easily identify if this is the correct answer by simply knowing that at least two of the options are correct. Similarly

if they could know that at least one option is not correct, they can exclude this choice. AOTA can be a clue to the student when the stem is wrongly formulated to refer to a plural choice. The student may read the first option, determine that it is correct and could be misled into choosing it without reading all the options.

Complex forms

The joined options appear in many combination formats. Examples are A and B or A and B but not C. This pattern of complex options is used only in K-type of MCQs which is not applied currently. This complexity makes the question more difficult or more confusing for some students as it causes heavy mental load. In addition, research has shown that it is less discriminating than items with distinct options.

NOTA (None of the above)

NOTA measures students' ability to detect the incorrect answers. If NOTA was the correct choice, we must be certain that there are no other options with any degree of correctness. If we are using choose the correct answer format (which is no longer preferred), NOTA can be used in some minimal situations, however, it is better to be reformulated. For example, if we are asking about the likely treatment, it can be reworded into no medication is needed instead of NOTA. The computer shuffling is another issue that needs to be considered as this choice can appear as option A; hence, the word "none of the above" is not suitable. Besides, if the stem was negative, NOTA will represent double negative and is not recommended.

It is to be noted that if we are using the recommended SBA type, none of the above is not suitable; because NOTA is either the key (this is wrong, other options are partially correct) or it is not the key and excluded from the start (7–10, 19, 21–28, 32).

Avoid vague terms (79%)

As stated previously in Item Content – avoid tricky items and The MCQ Stem – avoid absolute terms, these terms lack precision and there is seldom agreement on their actual meaning. A long list of words such as absolutely, rarely, scarcely, sometimes, usually, often or frequently. No actual interpretation of these terms and they can confuse examinees. These terms are not recommended in the question and stem; meanwhile, they are prohibited in the options (7, 10, 19, 21–22, 24–28).

Avoid test-wise item flaws (items that can be answered without studying) (79%)

Clues

As described previously in Item Format – avoid clueing, cross-referring or hinging, clueing means the presence of relational point in the syntax or the grammar, which may reveal an inference to the answer. The correct option is more likely to flow grammatically or by syntax from the item stem which can cue examinees to the correct answer.

Repetition

Similar wording or word repeats in the stem or in the options allows students to identify the correct option, without studying the contents of the course. This can be either word or phrase. Sometimes, a word is repeated by the meaning or inference without mentioning the word itself, e.g., a stem speaks about menstrual condition and one of the options comes with a term like amenorrhea.

Convergence

Question writers tend to use the correct answers more frequently across all options, and the testwise students will identify this as the correct answer. Students can often guess the answer by eliminating one incorrect

response and all options containing this response, or by selecting the responses that appear most frequently in all the options (convergence strategy). An example for this, when an item writer asked about the action of certain muscles. He wrote the options as following: flex and adduct, flex and abduct, flex and lateral rotate, extend and adduct. The student here will count and select the flex and adduct as they were the more repeated. In many situations, this type of wiseness can specify the correct answer. If we cannot avoid these patterns of questions, it is necessary to put all combinations and all doubles equally. In addition, if there are options with word like increased or decreased, we need to be sure that equal number of options having either words. This is also applied in all options which contain similar words, or phrases, make sure that number of these similar terms or phrases is balanced (7–10, 19, 21–22, 24, 26–28, 32).

Finally, if we need to look at assessment in depth; with many regulations, bylaws, and policies, the work force or faculty member factor is very crucial. Many of them are clinicians or busy in some aspects of the medical care. The nomenclature and language nature of the detailed guidelines may decrease compliance to them. Biased or flawed assessment items may be dangerous for the exam validity, reliability and this might make the assessment questionable; with further consequences on the future physician's abilities.

CONCLUSION

Compilation of MCQs guidelines was possible by simplification and word reduction to produce a shortlist of guidelines and a novel coloured pocket card. The card design is suitable as a pocket guide for the busy faculty. Adoption of this list and the coloured card might be a good asset for increasing compliance, improving the item-writing quality, enhancing the assessment efficiency and ensuring valid student results.

REFERENCES

1. Scouller K. The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay. Higher Education. 1998;35(4):453–72. https://doi.org/10.1023/A:1003196224280

2. van der Vleuten CP, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. Best Pract Res Clin Obstet Gynaecol. 2010;24(6):703–19. https://doi.org/10.1016/j.bpobgyn.2010.04.001

3. Wadi M, Abdul Rahim AF, Yusoff MSB, Baharuddin KA. The effect of MCQ vetting on students' examination performance. Education in Medicine Journal. 2014;6(2):e16–26. https://doi.org/10.5959/eimj.v6i2.216.

4. Downing SM. 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. Adv Health Sci Educ. 2005;10(2):133–43. https://doi.org/10.1007/s10459-004-4019-5

5. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ. 2004;38(9):974–9. https://doi.org/10.1111/j.1365-2929.2004.01916.x

6. Al-Faris EA, Alorainy IA, Abdel-Hameed AA, Al-Rukban MO. A practical discussion to avoid common pitfalls when constructing multiple choice questions items. J Fam Community Med. 2010;17(2):96–102. https://doi.org/10.4103/1319-1683.71992

7. Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education. 1989;2(1):37–50. https://doi.org/10.1207/s15324818ame0201_3

8. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education. 2002;15(3):309–33. https://doi.org/10.1207/S15324818AME1503_5

9. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. 3rd ed. Philadelphia, PA: National Board of Medical Examiners® (NBME®); 2003 [cited 4 September 2016]. Available from: https://www2.med.wayne.edu/elab/guides/exam_writing/NBME_Test_Question_Writing_Guide.pdf

10. Paniagua M, Swygert K. Constructing written test questions for the basic and clinical sciences. 4th ed. Philadelphia, PA: National Board of Medical Examiners® (NBME®); 2016 [cited 20 February 2019]. Available from: https://www.nbme.org/publications/item-writing-manual.html

11. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med. 2002;77(2):156–61. https://doi.org/10.1097/00001888-200202000-00016

12. Balaha M, El-Baramawi M, El-Hawary E. Three option multiple choice questions had the least non-functioning distractors: analysis of 1855 MCQs in first year competency based medical program at Tanta Faculty of Medicine, Egypt. International Journal of Scientific and Engineering Research. 2019;10(2):1432–8. Available from: https://www.researchgate.net/publication/331565259

13. Messick S. Test validity: a matter of consequence. Social Indicators Research. 1998;45(1):35–44. https://doi.org/10.1023/A:1006964925094

14. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830–7. https://doi.org/10.1046/j.1365-2923.2003.01594.x

15. Hassan S, Hod R. Use of item analysis to improve the quality of single best answer multiple choice question in summative assessment of undergraduate medical students in Malaysia. Education in Medicine Journal. 2017;9(3):33–43. https://doi.org/10.21315/eimj2017.9.3.4.

16. Ellsworth RA, Dunnell P, Duell OK. Multiple-choice test items: what are textbook authors telling teachers? The Journal of Educational Research. 1990;83(5):289–93. https://doi.org/10.1080/00220671.1990.10885972

17. Richman H, Hrezo M. The trouble with test banks. Perspectives in Learning. 2017;16(1):289–93.

18. Egyptian Knowledge Bank. 2019 [cited February 2019]. Available from: https://www.ekb.eg/en/web/researchers/home.

19. Collins J. Writing multiple-choice questions for continuing medical education activities and self-assessment modules. RadioGraphics. 2006;26(2):543–51. https://doi.org/10.1148/rg.262055145

20. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Educ Today. 2006;26(8):662–71. https://doi.org/10.1016/j.nedt.2006.07.006

21. Wood T, Cole G. Developing multiple-choice questions for the RCPSC certification examinations. Canada: Royal College of Physicians and Surgeons of Canada; 2004 [cited 1 April 2019]. Available from: https://www.scribd.com/document/142391188/Developing-Mcqs-for-RCPSC

22. Saudi Commission for Health Specialties. Item writing manual for multiple-choice questions. Item Writing and Style Guide 1st ed.; 2011 [cited 29 March 2019]. Available from: https://www.scfhs.org.sa/MESPS/PME/HowPrepareExams/MCQMethod/Documents/MCQ_Manual_Style_Guide_1stEdition[2]_FINAL.pdf

23. Brame CJ. Writing good multiple-choice test questions: writing good multiple-choice test questions. Nashville, TN: Vanderbilt Center for Teaching; 2013.

24. Schuwirth L, Pearce J. Determining the quality of assessment items in collaborations: aspects to discuss to reach agreement. Australia: Australian Medical Assessment Collaboration (AMAC); 2014 [cited 2 April 2019]. https://www.acer.org/files/quality-determination-of-assessment-items-amac-resource.pdf

25. Tenore A, Mathysen DGP, Mills P, Westwood M, Rouffet J-B, Papalois V, et al. A guide to successfully writing MCQs: executive summary. Belgium: European Council for Specialist Medical Assessments; 2015.

26. Davis MH, Ponnamperuma G, McAleer S, Rowley D. Joint committee on intercollegiate examinations: examiners' manual. Scotland, UK: University of Dundee; 2016 [cited 1 April 2019]. https://www.jcie.org.uk/documents/examiners_manual.pdf

27. Zimmaro DM. Writing good multiple-choice exam questions. Austin, TX: University of Texas; 2016 [cited 13 March 2019]. https://facultyinnovate.utexas.edu/sites/default/files/writing-good-multiple-choice-exams-fic-120116.pdf

28. Center for Leadership and Innovation in Medical Education (CLIME). Writing multiple choice questions guidelines and examples. University of Washington, School of Medicine; 2017. http://clime.washington.edu/resources-teaching/2017/7/6/preparing-test-items

29. Tarrant M, Ware J. A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. Nurse Educ Today. 2010;30(6):539–43. https://doi.org/10.1016/j.nedt.2009.11.002

30. Crooks, Terence J. The impact of classroom evaluation practices on students. Review of Educational Research. 1988;58(4):438–81. https://doi.org/10.3102/00346543058004438

31. 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. https://doi.org/10.1080/01421590802155597

32. Rodriguez MC. The art and science of item writing: a meta-analysis of multiple-choice item format effects. Annual meeting of the American Educational Research Association; April 1997 [cited 29 March 2019]; Chicago, IL. Available from: http://edmeasurement.net/research/artandscience.pdf