Article published online: 2022-08-31

Construction and Writing Flaws of the Multiple-Choice Questions in the Published Test Banks of Obstetrics and Gynecology: Adoption, Caution, or Mitigation?

Magdy H. Balaha[1,2] Mona T. El-Ibiary[2] Ayman A. El-Dorf[2] Shereef L. El-Shewaikh[2] Hossam M. Balaha[3]

[1] Department of Obstetrics and Gynecology, College of Medicine, King Faisal University, Kingdom of Saudi Arabia
[2] Department of Obstetrics and Gynecology, Faculty of Medicine, Tanta University, Tanta, Egypt
[3] Department of Computer Engineering and Systems, Faculty of Engineering, Mansoura University, Mansoura, Egypt

Address for correspondence Magdy H. Balaha, Department of Obstetrics and Gynecology, Faculty of Medicine, Tanta University, Al Geish St. Medical Campus, Tanta 31527, Egypt (e-mail: magdy.balaha@med.tanta.edu.eg; abcd221221@gmail.com).

Avicenna J Med 2022;12:138–147.

Abstract

Background The item-writing flaws (IWFs) in multiple-choice questions (MCQs) can affect test validity. The purpose of this study was to explore the IWFs in the published resources, estimate their frequency and pattern, rank, and compare the current study resources, and propose a possible impact for teachers and test writers.

Methods This cross-sectional study was conducted from September 2017 to December 2020. MCQs from the published MCQ books in Obstetrics and Gynecology was the target resources. They were stratified into four clusters (study-book related, review books, self-assessment books, and online-shared test banks). The sample size was estimated and 2,300 out of 11,195 eligible MCQs were randomly selected. The MCQs (items) were judged on a 20-element compiled checklist that is organized under three sections as follows: (1) structural flaws (seven elements), (2) test-wiseness flaws (five elements), and (3) irrelevant difficulty flaws (eight elements). Rating was done dichotomously, 0 = violating and 1 = not violating. Item flaws were recorded and analyzed using the Excel spreadsheets and IBM SPSS.

Results Twenty three percent of the items ($n = 537$) were free from any violations, whereas 30% ($n = 690$) contained one violation, and 47% ($n = 1073$) contained more than one violation. The most commonly reported IWFs were "Options are Not in Order (61%)." The best questions with the least flaws (75th percentiles) were obtained from the self-assessment books followed by study-related MCQ books. The average scores of good-quality items in the cluster of self-assessment books were significantly higher than other book clusters.

Conclusion There were variable presentations and percentages of item violations. Lower quality questions were observed in review-related MCQ books and the online-shared test banks. Using questions from these resources needs a caution or avoidance

Keywords
- multiple-choice questions writing flaws
- multiple-choice questions violations
- test bank quality

published online
August 31, 2022

DOI https://doi.org/10.1055/s-0042-1755332.
ISSN 2231-0770.

© 2022. Syrian American Medical Society. All rights reserved.
This is an open access article published by Thieme under the terms of the Creative Commons Attribution-NonDerivative-NonCommercial-License, permitting copying and reproduction so long as the original work is given appropriate credit. Contents may not be used for commercial purposes, or adapted, remixed, transformed or built upon. (https://creativecommons.org/licenses/by-nc-nd/4.0/)

Thieme Medical and Scientific Publishers Pvt. Ltd., A-12, 2nd Floor, Sector 2, Noida-201301 UP, India

strategy. Relative higher quality questions were reported for the self-assessment followed by the study-related MCQ books. An adoption strategy may be applied with mitigation if needed.

Introduction

There are several types of resources which contain multiple-choice questions (MCQs). These include study books, review books, self-assessment books, and online-shared question banks. These resources were created to assist students in preparing for specific topic content. The teachers and assessors in medical education may use the published MCQs in preparing their local course tests.[1,2] Students may use these resources and MCQs to be prepared for their examinations.

Valid and reliable testing represents a pledge that the student results are true and the actual performance is credible. The prevalence of item-writing flaws (IWFs) can artificially inflate or deflate student achievement, independent of their understanding of the content. As a result, the test construct validity is lowered. This represents a danger to the test validity and reliability.[3–5] Therefore, some inquiries arise about student performance and the validity of student achievement on these tests may be questionable.[6,7]

The MCQ guidelines pave the way for valid tests, content, cognitive levels, and item construction quality. Instructors should consider the cultural perspective to prepare correctly answerable items by the knowledgeable students, and without flaws that might help the nonknowledgeable ones.[8–12]

The evidence-based MCQ writing guidelines specified the item flaws as related to the structure, test wiseness, or the irrelevant difficulty.[4,11,13] Unnecessary verbal complexity, grammatical faults, and a lack of clarity or uniformity in the terminology are all examples of structural bias.[14] Test wiseness allows examinees to guess or deduce correct answers without knowing the material, thereby inflating their test scores. This gives an unfair advantage for the unprepared students who can guess over students who are not "test wise." Irrelevant difficulty makes a question difficult to be understood, for reasons unrelated to the content. This prevents prepared students from demonstrating their knowledge.[11,12,14]

Therefore, these MCQs in the published books and test banks need to be screened for item flaws which may affect the test validity and reliability. A review of the literature, documented the magnitude of item flaws in the published test banks, in psychology, business, law, and Nursing disciplines.[15–17] However, it did not reveal any published research in medical educational disciplines.

The scarcity of publications related to this issue may highlight an unspoken area in medical education. The current research is conducted on the published question review books or test banks in obstetrics and gynecology. Due to the similar curricular contexts, the findings of this research may have a broader impact to be generalizable to the other medical specialties.

The purpose of this study was to (1) explore the published resources for IWFs; (2) estimate their frequency and pattern in the different resources and compare them with other published studies; (3) rank and compare the current study resources; and (4) propose a possible impact for teachers and test writers.

Methods

Settings

An observational cross-sectional design was used, throughout the period from September 2017 to August 2020. Analysis was finished in December 2021. The available test bank books or resources to the Faculty members (test makers) and the students were from the near bookstores or the service from Egyptian online library (*www.ekb.eg*). The resources were mainly for undergraduate students with some for postgraduate students. The resources were stratified into four clusters as follows: cluster 1 (C1), study book–related MCQs with a companion MCQs section or separate MCQ book; cluster 2 (C2), review books were published to revise with samples of MCQs; cluster 3 (C3), self-assessment books were published as a review series to prepare for certain examinations; and cluster 4 (C4), the online-shared test banks included some personal or web site collective efforts to present questions for academicians.

Random Inclusion of Multiple-Choice Questions

We used Levy's sample size formula ($n = Z^2pq \,/\, e^2$), where n = minimum sample size, $Z = 1.96$ at 95% confidence interval, p = estimated prevalence of the event from previous literature, $q = (1-p)$, and e = margin of error (0.02). Here, n was calculated to be 2,285. A random item generator using Microsoft Excel was used. The randomly generated item numbers were highlighted in each resource and were plotted in the excel spreadsheet. Only the MCQs type was considered for this study. The resources, which contained at least 100 MCQs, were eligible for inclusion. The selected MCQ samples were of either 50, 100, 150, 200 or 250 items if the resources contained <200, "200–500," "500–1,000," "1,000–1,500," or >1,500 items, respectively. A total of 11,195 items from 20 resources were eligible (►**Supplementary Table S1**; available in the online version). A randomly selected 2,300 MCQs (20.5%) were selected for the review process.

Item Review Checklist

A 20-element checklist was adapted after reviewing the literature.[11,16,18–20] The checklist elements were allocated under three sections: structural flaws (seven), test-wiseness flaws (five), and extra irrelevant difficulty flaws (eight). The checklist elements were rated on a dichotomous scale, 0 = violating and 1 = not violating. An item was classified as

 © 2022. Syrian American Medical Society. All rights reserved.

flawed if it violated ≥1 of these elements (►Supplementary Table S2; available in the online version).

After data entry to the spreadsheet, it becomes visible for each reviewed question, how many points take "1" that represent points of no violation. This was calculated for each of the three sub-categories, structural, test wiseness, and irrelevant difficulty, as well as the total item flaws. If absent flaws or violations in all of them, it is a good item.

The content and cognitive level of the MCQs were not included in the checklist due to two reasons. First, these resources represented variable academic targets. Second, they represented a self-assessment for some courses at certain contexts, with variable cognitive levels.

Reliability of the Checklist

The checklist consistency was evaluated by the pilot group of MCQs, which comprised 50 MCQs that were chosen from five different books. The elements showed either violation, or no violation. Reliability was acceptable (α Cronbach = 0.752). The items that were used in the pilot study were not included in the final study.

Raters of Multiple-Choice Questions

This review was conducted with multiple raters who were practicing joint review sessions till reaching an acceptable level of interrater agreement. The first author was working at the King Faisal University, KSA (2007–2016), where he was trained in educational and assessment strategies at the University of Groningen, the Netherlands, during a process of new medical curriculum reform.[21] Multiple Faculty development sessions were conducted at both our current Colleges, as well as at the Gynecology Department levels, where assessment and MCQs were trained on. The second, third, and fourth authors attend these workshops, and they practiced on the item writing guidelines and used the made checklist in the review process. These workshops was accredited from the American Academy of medical education. The last author, being an expert in informatics, prepared the materials for review from different resources, compiled the files in Excel and SPSS.

Interrater Reliability (Concordance)

The included resources were randomly distributed in a balanced way among the researchers to avoid any preferential bias. Three joint sessions of assessment of some MCQs were done followed by discussion. The activity was repeated twice in each of the three sessions to reach a joint judgement. After that, a session for concordance testing was done. Ten MCQs were chosen for independent evaluation. Kappa (k) coefficient was calculated. This was tried twice until the required coefficient of the agreement was reached ($k = 0.82$). Lastly, final reconciling of the scoring variance was done. All piloted questions were not included in the final study.

Review Process

Each resource was read twice. First, aimed for screening of the flaws. Second, the flaws were analyzed according to the checklist. The screening of 50 MCQs required approximately 10 hours. The first set of books, containing <200 MCQ, was finished in a 6-month span, while the sources containing ≥1,500 MCQs needed a 10-month span. The reviewing was accomplished at the researchers' convenience. The time spent in the analysis and rating was 30 minutes, whereas the subsequent review consumed 10 to 15 minutes for each MCQ. The item flaws were recorded and computed in a premade Microsoft Excel spreadsheet.

Statistical Procedures

The final compiled master spreadsheet was exported to SPSS (IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY; IBM Corp) for descriptive and inferential statistical analysis.[22] The frequency and cross-tabulation were computed. The scores were calculated as counts and percentage of violations per every question (item) and among the whole resources. After calculation per total reviewed questions, again, this analysis was done at the level of book clusters as mentioned above in the resource classification as Ca, C2, C3, and C4. The 25th, 50th, and 75th percentiles were generated for the different resources' violation frequencies. Comparing the violations in the different resources was done using the nonparametric analysis of variance "Kruskal–Wallis H-test."[23] Significance was considered at "$p = 0.05$."

Results

The good items, "free from any violations," were seen in 23% ($n = 537$), whereas 30% ($n = 690$) contained one violation, and 47% ($n = 1073$) contained more than one violation. The average percentages of items with no structural, test wiseness, and irrelevant difficulty violation were 62, 87, and 36%, respectively (►Table 1).

The distribution of item violations were presented under each section of the checklist. The most prevalent violations was "the options are not arranged in order (60%)." Violations in nine parameters ranged from 10 to 18%. The violations in the remaining 10 points of the checklist were less than 10% (►Table 2).

The different percentile values (25th, 50th, and 75th) for the overall, structural, and irrelevant difficulty violations were determined after ranking of all resources. Since the test-wiseness violations were minimal as seen in ►Table 1, they were not included in the analysis. Values >75th percentiles for the total violations, as well as the irrelevant difficulty violations, were present in self-assessment books. Values >75th percentiles for the structural violations were obtained from self-assessment and study-related books. The questions with the least percentile (contain most flaws) were obtained from the online shared resources (►Table 3).

The average percentages of the items without overall, structural, test wiseness, and irrelevant difficulty violations in the self-assessment cluster of books were 38, 66, 90, and 54%, respectively. These values were higher compared with other clusters as shown by the Kruskal–Wallis H-test. There was a statistically and highly significant difference in the overall item ($\chi^2 = 190.54$), structural ($\chi^2 = 92.4$), test wiseness

 © 2022. Syrian American Medical Society. All rights reserved.

Table 1 Distribution of the percentage of items having an overall item flaws (violations), structural, test wiseness, and irrelevant difficulty flaws in all MCQ resources

	Total item flaws (%)			Structural flaws (%)			Test-wiseness flaws (%)			Irrelevant difficulty flaws (%)		
Source or book[a]	Good items[b] ($n=537$)	One flaw ($n=690$)	>1 flaw ($n=1,073$)	No flaws ($n=537$)	One flaw ($n=690$)	>1 flaw ($n=1,073$)	No flaws ($n=537$)	One flaw ($n=690$)	>1 flaw ($n=1,073$)	No flaws ($n=537$)	One flaw ($n=690$)	>1 flaw ($n=1,073$)
1	10	14	76	42	8	50	68	28	4	10	64	26
2	4	16	80	38	6	56	70	28	2	8	68	24
3	12	28	60	62	28	10	80	18	2	26	54	20
4	14	30	56	50	46	4	78	20	2	34	26	40
5	4	18	78	46	34	20	60	32	8	12	34	54
6	11	29	60	37	34	29	81	18	1	40	36	24
7	4	44	52	58	42	0	84	14	2	8	76	16
8	36	19	45	51	17	32	90	9	1	70	26	4
9	52	23	25	63	35	2	82	16	2	84	15	1
10	68	19	13	78	21	1	97	3	0	81	18	1
11	43	27	30	75	11	14	92	6	2	53	38	9
12	23	21	56	49	31	20	90	8	2	38	38	24
13	45	17	38	74	21	5	94	6	0	49	31	20
14	8	18	74	44	36	20	90	10	0	12	44	44
15	16	21	63	57	22	21	91	9	1	24	36	40
16	7	15	77	39	27	34	85	14	1	12	51	37
17	34	42	24	77	18	5	89	9	2	50	48	3
18	13	59	28	83	16	2	89	11	1	17	77	7
19	15	35	50	65	22	14	89	10	1	29	44	27
20	24	39	37	76	14	10	93	6	1	35	44	20
Average	23	30	47	62	23	15	87	11	1	36	44	20

Abbreviation: MCQ, multiple-choice question.
[a]As denoting to the selected book and/or resource (List of sources was detailed in **Supplementary Material S1** [available in the online version]).
[b]Good items mean Items without any flaw (violation).

Table 2 The differential distribution of different violations and their percentage out of the 1,763 MCQs, (1,763/2,300 = 77%) which contain item flaws (537 good items, i.e., without flaws)

Violation criteria		Violations per 1,763 MCQs	% (NB)
V No.	Structural violations (flaws)		
V1	The item is not conclusive	123	7
V2	The item is not typical SBA type	280	16
V3	The item is not focused	310	18
V4	The item is not clearly expressed	251	14
V5	All the options are not uniform	260	15
V6	All the options are not homogenous	230	13
V7	All the options are not plausible	108	6
	Test-wiseness violations		
V8	There is clang	30	2
V9	There is clueing	72	4
V10	There is convergence	14	1
V11	There is absolute or vague terms	85	5
V12	There is crowded key option	122	10
	Irrelevant difficulty violations		
V13	The item is overloaded with information	169	10
V14	The item is not stated positively	250	14
V15	The options are not arranged in order	1,067	61
V16	The options have numerical Inconsistency	9	1
V17	The options have overlap	75	4
V18	The options have ambiguity	134	8
V19	There is “all of the above” or “none of the above”	228	13
V20	The options have “complex choices”	71	4

Abbreviations: MCQ, multiple-choice question; NB, the major violations in the current study were compared with the different resources and was presented in **Supplementary Material S3** (available in the online version); SBA, single-best answer; V, violation.
Note: Violation no. in the checklist (the detailed checklist was mentioned in the **Supplementary Material S2** [available in the online version]).

($\chi^2 = 36$), and irrelevant difficulty violations ($\chi^2 = 237$), respectively ($p \leq 0.01$; ►**Table 4**).

►**Table 5** explore the IWFs in the current, as well as other published educational literature. The number of studied items in the current study was higher than most of the other presented studies. The violation rates ranged from 60 to 77%, with the highest rate in the current study, that is, 77%. The percentage of good nonviolated items was 23% which was the lowest if compared with other studies.

Discussion

Despite the availability of many item-writing guidelines, IWFs are common in MCQs. The quality of these MCQs is of paramount importance, as they would affect the validity and reliability of the examinations.[3–7] Downing reported that because of flawed MCQs, as many as 10 to 15% of students could fail a test they should have passed.[14]

The first purpose of this study was to explore the published resources for possible IWFs. In the current study, the good items which are free from any violations were seen in 23% of the reviewed MCQs. Also, 30% contained one, whereas 47% contained more than one violations. The average percentages of items with no structural, test wiseness, and irrelevant difficulty violations (flaws) were presented in ►**Table 1** and showed that structural and irrelevant difficulty violations were more visible in the current study.

To the knowledge of the author, there were no published studies about the quality of MCQs derived from published books in medical disciplines. The authors reviewed the IWFs in other published educational literature. These publications' findings were summarized in ►**Table 5**.[13,15–17,26–29] The overall view as seen in ►**Table 5**, projected a common phenomenon of the prevalence of IWFs in the test banks. All the studies dispose of a common pattern of increased item violation percentage, excluding Bailey et al.[27]

Tarrant et al collected 2,770 MCQs from one nursing department. Also, 14.1% of these MCQs were teacher generated, 36.2% were taken from test banks, and 49.4% had no identified source. They show that nearly 54% of the questions contained IWFs. One violation was present in 34%, while in 20% of MCQs, more than one violation was present. Tarrant

 © 2022. Syrian American Medical Society. All rights reserved.

Table 3 Distribution of the violations according to the different percentile values for the overall violations, as well as, the structural and irrelevant difficulty violations (ranking of the resources was done with sources with the least violations were at the top)

Percentile	Source[a]	No_Viol (%)	Percentile	Source[a]	No_S_Viol (%)	Percentile	Source[a]	No_D_Viol (%)
>75th (>35.5)	10	68	>75th (>74.8	18	83	>75th (>49.5)	9	84
	9	52		10	78		10	81
	13	45		17	77		8	70
	11	43		20	76		11	53
	8	36		11	75		17	50
50th–75th (15–35.5)	17	34	50th–75th (>57.5–74.8)	13	74	50th–75th (>31.5–49.8)	13	49
	20	24		19	65		6	40
	12	23		9	63		12	38
	15	16		3	62		20	35
	19	15		7	58		4	34
25th–50th (8.5–14.5)	4	14	25th–50th (44.5–57.5)	15	57	25th–50th (8.5–14.5)	19	29
	18	13		8	51		3	26
	3	12		4	50		15	24
	6	11		12	49		18	17
	1	10		5	46		5	12
<25th (<8.5)	14	8	<25th (<44.5)	14	44	<25th (<12)	14	12
	16	7		1	42		16	12
	2	4		16	39		1	10
	5	4		2	38		2	8
	7	4		6	37		7	8

No_D_Viol, items without irrelevant difficulty flaws; No_S_Viol, items without structural flaws; No_Viol, total good items without flaws.
Note: Test-wiseness violations were not included as they were minimal as seen in ►**Table 1**.
[a]Sources: • Study-book related multiple-choice questions (MCQs) are resources Coded: 1, 2, 3, 18, 19, and 20. • Review-book related MCQs are resources Coded: 4, 5, and 6. • Self-assessment books are resources Coded: 8, 9, 10, 11, 12, 13, 15, and 17. • Online shared MCQs are resources Coded: 7, 14, and 16.

 © 2022. Syrian American Medical Society. All rights reserved.

Table 4 Statistical analysis of the differential distribution the percentage of item without violations (flaws) after clustering of the resources into C1, C2, C3, and C4

Overall items violations	Good items[a] (%)	One flaw (%)	>1 flaw (%)	Structural violations	No flaws (%)	One flaw (%)	>1 flaw (%)
C1: study book–related MCQs	16	39	45	C1: study book–related MCQs	69	17	14
C2: review book–related MCQs	10	26	64	C2: review book–related MCQs	42	37	21
C3: self-assessment books[b]	38	25	37	C3: self-assessment books[b]	66	22	12
C4: online-shared MCQs	7	21	72	C4: online-shared MCQs	44	32	24
χ^2 (KW) = 190.54, $p = 0.000$				χ^2 (KW) = 92.4, $p = 0.000$			
Test-wiseness violations	No flaws (%)	One flaw (%)	>1 flaw (%)	Irrelevant difficulty violations	No flaws (%)	One flaw (%)	>1 flaw (%)
C1: study book–related MCQs	87	12	1	C1: study book–related MCQs	25	55	20
C2: review book–related MCQs	75	22	3	C2: review book–related MCQs	32	33	35
C3: self-assessment books[b]	90	9	1	C3: self-assessment books[b]	54	33	13
C4: online-shared MCQs	86	13	1	C4: online-shared MCQs	11	53	36
χ^2 (KW) = 36, $p = 0.000$				χ^2 (KW) = 237, $p = 0.014$			

Abbreviations: MCQ, multiple-choice question.
Note: χ^2 (KW) Kruskal–Wallis *H*-test.
[a]Items without any flaw (violation).
[b]Highly significant ($p < 0.01$).

Table 5 [b]Comparison of the number of the reviewed questions and percentage of different items in the previously published, with the currently reported results

No.	Reference	Year	Discipline	No of test banks or books	Reviewed questions	Good items[a]	One-item flaws	>1 item flaws
1	The current study	2022	Medical discipline	20	2,300	23	30	47
2	Ellsworth et al[15]	1990	Educational psychology	14	1,080	39	44	17
3	Hansen and Dexter[16]	1997	Business auditing	10	400	25	42	33
4	Garrison et al[26]	1997	Business law	11	440	33	46	21
5	Bailey et al[27]	1998	Accounting	16	100	94	6	
6	Masters et al[17]	2001	Nursing	17	2,913	24	76	
7	Moncada and Harmon[28]	2004	Accounting	5	684			
8	Tarrant et al[13]	2006	Nursing	–	997[c]	46	34	20
9	Ibbett and Wheldon[29]	2016	Financial accounting	6	263	33	56	10

[a]Good items mean Items without any flaw (violation).
[b]This table is not one of our results, however, it is a review of different resourced that would serve comparison in the discussion.
[c]2,770 multiple-choice questions were reviewed, out of them 36% (997) were derived from test-banks.

et al claimed the high proportion of flawed questions to be affected by test banks.[13]

This trend of broad concordance of the current study findings with other test banks–related studies might support the generalization of the current study findings in other medical disciplines and increase the external validity; however, this may need extra evidence.

The second purpose of this study was to estimate the frequency, and pattern of IWFs in the different resources, and compare them with other published studies. The findings of the current study was presented in ►Table 2. The distribution of IWFs in the other research describing the test banks in different disciplines was thoroughly reviewed (►Supplementary Table S3; available in the online version).[13,15,16,26,29] The description of IWFs in different studies showed variable presentations. IWFs were presented as per the employed guidelines in each research which varied from a 9-item to a full exhaustive 31-item checklist. Moreover, the

 © 2022. Syrian American Medical Society. All rights reserved.

wording differs from being stated as a guideline for the good item or as points to be avoided. Some of the researchers calculated the percentage of IWFs out of the total reviewed MCQs; hence, decreasing the prevalence. In the current study, adopts calculating the IWFs as a percentage of the flawed items.

Masters found 2,233 IWF s in 2,913 questions in test banks with the nursing textbooks, without detailing the prevalence of the different types.[17] The current study showed a very high percentage (61%) of violated role no. 15 in the checklist, that is, "the options must show alphabetical, logical, or numerical order." This rule was only reported by Ibbett and Wheldon as 31%.[29] This simple rule need minimal effort to edit the MCQ to avoid or correct the violation.

The demonstrated violations (IWFs) in the current study were either simple and can be easily corrected with little or minimal mitigation efforts or advanced that may need extra educational skills and caution for its modification or correction. Simple IWFs included "options are not in order (61%)," "options are not uniform (15%)," "options are not homogenous (13%)," and "crowded key option (10%)." Advanced IWFs included "item is not focused (18%)," "item is not typical single-best-answer (SBA) type (16%)," "item is not clear (14%)," "item is overloaded (10%)," "item is not positively stated (14%)," and "all of the above to none of the above (13%)."

The percentages of violations per different resources were calculated and presented in ►Table 5, with a common viewed pattern of increased item violation percentage, excluding Bailey et al.[27] The violation rates ranged from 60 to 77%, with the highest rate in the current study. The percentage of good nonviolated items was 23% which was the lowest if compared with other studies.

The third purpose of this study was to rank and compare the different resources in the current study. The percentages of the items without violations were ranked and sorted with the calculation of the 25th, 50th, and 75th percentiles, where the sources that have questions with the least number of flaws (best items) are listed at the top of the table (75th percentiles) and the resources with greater number of flawed questions are at the bottom of the table.

From the values >75th and 25th percentiles for the total, structural, and irrelevant difficulty violations, it was clearly evident that the resources in the 75th percentile were mostly self-assessment books (plus a couple of study-related books). The questions with the most flaws (25th percentiles) were obtained from the online-shared resources.

This led us to the idea of clustering of the studied resources as per their type whether study-related, review-related, or self-assessment books, as well as the online-shared test banks (►Table 4). The average scores of good items without any violations, as well as the items without structural and irrelevant difficulty violations, were significantly higher in the self-assessment cluster of books than the other book clusters. Moreover, the average scores of items with more than one violation in the online-shared test banks were significantly higher than the other book clusters. The current research included a unique feature of sorting the test banks according to the sources with fewer violation (self-assessment cluster of books) and others with the maximum violation (online-shared test banks). Nevertheless, it needs further studies in other disciplines to be accepted as a universal concept.

A possible explanation could be ascribed to the aim of these books or test banks. Self-assessment books were primarily published as a course review and as a prior step for certain examinations; therefore, they consider the item-writing guidelines. The study-related and the review-related books are linked to content covering specific topics. Their authors focused on the study content and not on the MCQ quality. The online-shared test banks were personal uploads or web site files that were published for revision. They might be just a collective effort of the memorable items from different resources and in most of them, without consideration of guidelines.

According to Tarrant et al, test banks are frequently made available to users as an incentive to utilize a textbook for the course, although textbook authors may not have formal training in MC item development or are not the people who develop the test bank items.[13]

The fourth purpose of this study was to propose a possible impact of the current study. From the findings as seen in ►Tables 3 and 4, the utilization of the published MCQs books with such percentage of IWFs for making the in-home examinations without adequate scrutiny could be a threatening source for the examination validity. So long as the examinations have a role in assigning student grades, credible student results would be also affected.

Tarrant and Ware evaluated 10 summative test papers in one nursing program, using a total actual examination scale and a standard scale of the unflawed items only. The proportion obtaining a score ≥80% was 20.9% on the standard scale, versus 14.5% on the total scale. Hence, high-achieving students were more likely to be penalized for flawed items.[24] Pham et al designed a cross-over study with 100 pairs of MCQs with and without IWFs. The mean item scores were positively impacted by "correct longest choice," "clues to the proper response," and "implausible distractors," while the mean item scores were negatively impacted by "central idea in options rather than stem." They concluded that IWFs produced errors that are neither systematic nor predictable and this unpredictability results in loss of examination validity.[25]

While teacher and item writers, due to many reasons, may use outsources,[1,2] they are essentially required to apply the guidelines in constructing questions for in-home examinations. They are required to have suitable training with a perspective on the quality of item writing, postexamination validation to ensure reasonable content validity, as well as construct validity.[30,31] Some researchers proved that the locally prepared items for the in-home examinations were better than MCQs taken from the published test banks.[32,33]

Some of the demonstrated IWFs in the current study were simple and could be easily corrected. The effort that could be done to improve the quality of "in-house" medical school examinations might rely on the vigilant approach to detect

 © 2022. Syrian American Medical Society. All rights reserved.

the IWFs. Some of the violations may need an extra effort for mitigation or correction and rewording of the questions by the instructors and teachers. By this, the quality of the MCQs may be guaranteed by reducing or correcting the IWFs.

Besides the ethical and copyright concerns, when using the published MCQ resources, teacher and item writers are required to have a visionary approach before dealing with any published MCQs resources. Some resources as the self-assessment cluster followed by the study review books have a good percentage of items without violations (flaws), while the online shared resources did not have this advantage. Dealing with any resource needs a vigilant approach to clarify whether it follows the guidelines or not. Items without violations ensure examination validity.

Teachers and test writers might use a three-tier strategy to deal with items or MCQs according to the degree of their accuracy. This strategy includes the adoption (if free from violation), mitigation (to make all possible corrections before using) or caution (if having many violations).

Finally, as most of medical publications focus on content and technical areas, the current study was the first to evaluate such a quality domain or an unspoken area, so it might be a motive for further studies in different medical disciplines. The study was not planned to criticize these resources, meanwhile it aimed to highlight the flaws and propose an approach for their possible utilization.

Limitations

There were some limitations to this study. No similar previous publications in the field of medical disciplines, so the comparison and analysis of the results were limited to the publications in other disciplines. The current study was limited to "one medical discipline." Further research needs to include other medical disciplines to make the results more generalizable. Publishing this study may encourage others to share in the rating of this unspoken area or MCQ quality in the future. The teachers and students who use the MCQs were not interviewed, so it was difficult to survey either the teachers about the extent to which they create and/or use MCQs supplied by published test banks or the students for their dependence on and utility of these resources. Moreover, as this issue may be felt stigmatizing by some persons, so that the response may need some official central anonymous organized effort, coupled with its relation to the postexamination analysis to assess the MCQs psychometric properties.

Conclusion

The current study has documented, for the first time, the variable quality of the published MCQs in one medical discipline, "obstetrics and gynecology." The distribution of item violations was variable among the different clusters of published MCQs resources. A three-tier strategy to deal with such items was proposed.

The lower quality questions were observed in review-related MCQ books and the online-shared test banks. Using questions from these resources needs a caution strategy or avoidance to avoid unfair student assessment.

Relative higher quality questions were reported for the self-assessment followed by the study-related MCQ books which have higher percentages of items without violations. There were fewer items with flaws and most of them could be easily managed to improve the MCQ quality. An adoption strategy may be applied or mitigation if needed, prior to their adoption.

Authors' Contributions

M.H.B. contributed with idea, concept, design, acquisition of data, studying of collected data, data analysis, interpretation of data, drafting the article, revising the article, and shared final approval.

M.T.El.-I. took part in design, studying of collected data, interpretation of data, revising the article, and also shared final approval.

A.A.El.-D. took part in conceptualization, design, studying of collected data, interpretation of data, revising the article, and shared final approval.

S.L.El.-S. took part in conceptualization, design, studying of collected data, interpretation of data, drafting the article, revising the article, and shared final approval.

H.M.B. contributed with conceptualization, design, acquisition of data, data analysis, drafting the article, revising the article, and shared final approval.

Funding

None.

Conflict of Interest

None declared.

References

1 Farley JK. The multiple-choice test: writing the questions. Nurse Educ 1989;14(06):10–12, 39
2 Vyas R, Supe A. Multiple choice questions: a literature review on the optimal number of options. Natl Med J India 2008;21(03):130–133
3 Cynthia B, Whitney DR. The effect of selected poor item-writing practices on test difficulty, reliability and validity. J Educ Meas 1972;9(03):225–233
4 Downing SM. Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? Acad Med 2002;77(10(Suppl 10):S103–S104
5 Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003;37(09):830–837
6 Idika DO, Shogbesan YO, Bamikole OI. Effect of test item compromise and test item practice on validity of economics achievement test scores among secondary school students in cross river state, Nigeria. African Journal of Theory and Practice of Educational Assessment 2016;3(06):33–47
7 Breakall J, Randles C, Tasker R. Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. Chem Educ Res Pract 2019;20(02):369–382
8 Determining the quality of assessment items in collaborations: aspects to discuss to reach agreement. Developed by the Australian Medical Assessment Collaboration. Accessed July 19, 2022 at: https://www.acer.org/files/quality-determination-of-assessment-items-amac-resource.pdf
9 Tenore A, Mathysen DGP, Mills P, Westwood M, Rouffet J-B. Papalois V ea. UEMS-CESMA Guidelines. Accessed July 19, 2022

 © 2022. Syrian American Medical Society. All rights reserved.

at: https://www.uems.eu/news-and-events/news/news-more/uems-cesma-guidelines
10 Davis MH, Ponnamperuma G, McAleer SDR. The Joint Committee on Intercollegiate Examinations. Accessed July 19, 2022 at: https://www.jcie.org.uk/Content/content.aspx
11 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
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 (IJSER) 2019;10(02): 1432–1438
13 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(08):662–671
14 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 Theory Pract 2005;10(02):133–143
15 Ellsworth RA, Dunnell P, Duell OK. Multiple-choice test items: what are textbook authors telling teachers? J Educ Res 1990;83 (05):289–293
16 Hansen JD, Dexter L. Quality multiple-choice test questions: item-writing guidelines and an analysis of auditing testbanks. J Educ Bus 1997;73(02):94–97
17 Masters JC, Hulsmeyer BS, Pike ME, Leichty K, Miller MT, Verst AL. Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. J Nurs Educ 2001;40(01):25–32
18 Bosher S. Barriers to creating a more culturally diverse nursing profession. Linguistic bias in multiple-choice nursing exams. Nurs Educ Perspect 2003;24(01):25–34
19 Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality. Adv Health Sci Educ Theory Pract 2012;17(03):369–376
20 Nedeau-Cayo R, Laughlin D, Rus L, Hall J. Assessment of item-writing flaws in multiple-choice questions. J Nurses Prof Dev 2013;29(02):52–57, quiz E1–E2
21 Medical Sciences Board of Examiners. IBMG. Rules and Regulations for the BSc degree programme. Appendix 3–5. Faculty of Medical Sciences, University of Groningen, the Netherlands; 2011:25–32
22 Yockey RD. SPSS demystified: a simple guide and reference. London, United Kingdom: Routledge; 2018
23 Vargha A, Delaney HD. The Kruskal-Wallis test and stochastic homogeneity. J Educ Behav Stat 1998;23(02):170–192
24 Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 2008;42(02):198–206
25 Pham H, Besanko J, Devitt P. Examining the impact of specific types of item-writing flaws on student performance and psychometric properties of the multiple choice question. MedEdPublish 2018;7(04):4
26 Garrison MJ, Hansen JD, Knoepfle TW. An analysis of multiple-choice questions from business law testbanks and from the CPA examination. J Leg Stud Educ 1997;15(01):91–105
27 Bailey CD, Karcher JN, Clevenger B. A comparison of the quality of multiple-choice questions from CPA exams and textbook test banks. Accounting Educators' Journal. 1998;10(02):12–30
28 Moncada SM, Harmon M. Test item quality: an assessment of accounting test banks. Journal of Accounting & Finance Research. 2004;12(04):28–39
29 Ibbett NL, Wheldon BJ. The incidence of clueing in multiple choice testbank questions in accounting: some evidence from Australia. e-Journal of Business Education and Scholarship of Teaching 2016;10(01):20–35
30 Royal KD, Hedgpeth M-W, Posner LP. A simple methodology for discerning item construction flaws in health professions examinations. Health Prof Educ 2019;5(01):82–89
31 O'Neill LD, Mortensen S, Nørgaard C, Holm AL, Friis UG. Screening for technical flaws in multiple-choice items: a generalizability study. Dansk Universitetspaedagogisk Tidsskrift. 2019;14(26): 51–65
32 Richman H, Hrezo MJ. The trouble with test banks. Perspectives In Learning. 2017;16(01):3–8
33 Mulready-Shick J, Edward J, Sitthisongkram S. Developing local evidence about faculty written exam questions: Asian ESL nursing student perceptions about linguistic modification. Nurs Educ Perspect 2020;41(02):109–111

 © 2022. Syrian American Medical Society. All rights reserved.