Analysis of Multiple-Choice Questions In Musculoskeletal Module At orthopaedics and Traumatology Division, Syarif Hidayatullah State Islamic University Faculty of Medicine

Bisatyo Mardjikoen[1*], Ahmad Mu'taz Shobary[2], Achmad Zaki[3], Marita Fadhilah[4], Siti Nuraisyah Jauharo[4], Ahmad Azmul A. Irfan[6]

[1,3,6]Department of Surgery, Faculty of Medicine, Syarif Hidayatullah State Islamic University, Indonesia
[2]Undergraduate Medical Study Program, Faculty of Medicine, Syarif Hidayatullah State Islamic University, Indonesia
[4]Department of Community Medicine, Faculty of Medicine, Syarif Hidayatullah State Islamic University,Indonesia
[5]Department of Clinical Pathology, Faculty of Medicine, Syarif Hidayatullah State Islamic University,Indonesia
Corresponding author:
E-mail : *bisatyo@uinjkt.ac.id*

Abstract.
Background : Multiple choice question are a form of assessment to measure student competence. The quality of the multiple choice items can be determined using the parameters of the difficulty item, discrimination index, and distractor efficiency. Objective: To carry out item analysis of MCQs and the association between distractor efficiency with difficulty index and discrimination index in Musculoskeletal Module at Orthopaedic and Traumatology Division, Syarif Hidayatullah State Islamic University Faculty of Medicine, Jakarta, in 2018 - 2019. Methods: Observational analytic study with cross sectional design. Total samples analyzed were 155 items with 620 distractors. Results: The difficulty index showed 12 (7.75%) difficult items, 52 (33,55%) moderate items, and 91 (58,7%) easy items. Out of 155 items, 89 items (57.4%) were good questions and 66 items (42.6%) were poor questions, based on discrimination index. Eight items (5.2%) had high distractor efficiency, whereas the distractor efficiency of 73 items (47.1%), and 74 items (47.7%) were moderate and low, respectively. The functional distractors were 247 (39.8%) and non functional distractors were 373 (60.2%). There was a significant association between the distractor efficiency with difficulty index and discrimination index (p <0.05). Conclusion: There was a considerable number of nonideal items based on the difficulty index, discrimination index, and distractor effectiveness. The distractor efficiency was significantly associated with the difficulty index and discrimination index (p <0,05).

Keywords *: Difficulty Index, discrimination index, distractor efficiency, non functional distractor, multiple choice question and Musculoskeletal module.*

I. INTRODUCTION

Assessment is an important thing, especially in medical education. Gronlund (2009) says that the assessment is a systematic process of collecting, analyzing, and interpreting information or data to determine the extent to which students have achieved learning objectives [1]. Assessment is divided into several models according to Miller (1990) with the Miller Pyramid scheme. According to him, clinical competence is divided into four levels, namely: knows, knows how, shows, and does [2].

Multiple choice questions is one of the form of assessment. Multiple choice questions are questions with a limited response whose answer must be selected from several possible answers or options. This objective assessment instrument contains a stem or problem description, an introduction or question, and a list of answer options. Multiple choice questions have the most basic levels and the level above according to Miller Pyramid, namely knows and knows how [3] [4]. This type of question is a method of assessment in summative examinations at the Faculty of Medicine, UIN Syarif Hidayatullah Jakarta.

Good multiple choice questions must have good quality, too. Researchers assess the quality in several item analysis indicators, namely the difficulty index (DIFI), the discrimination index (DI or D), the distractor efficiency (DE), and internal consistency. The criteria for good quality questions have DIFI or p 0.30 - 0.70, DI ≥ 0.2, and DE 100%. Unqualified questions can have an impact on the bias of the assessment, because these questions cannot validly and reliably measure the extent to which students mastery of actual competence [5]. Several publications related to the analysis of multiple choice questions are research conducted by Rao from the Department of Pathology, KS Hegde Medical Academy, India, which found that 2 (5%), 34 (85%), and 4 (10%) items were easy, moderate, and difficult based on DIFI. Meanwhile, according to the DI parameter, 4 (10%) questions were categorized as good, 6 (15%) items were acceptable and 6 (15%) were categorized as poor. The question has 120 distractors which is 6 non functional and 114 functional.

Then, International Medical University, Malaysia, Burud found in the Department of Surgery and Obstetrics and Gynecology, there is 25 (20,8%), 35 (29.1%), 37 (44%), 13 (30.8%), and 10 (8.3%) items consecutively included very good, good, fair, poor, and negative discrimination based on DI. Meanwhile, there are NFD numbers in sequence 0, 1, 2, and 3 in 47 (13%), 51 (14%), 30 (8.3%), and 4 (1.1%) items. In addition, the analysis of multiple choice items carried out by Purwoko and Mundijo on the Genetic and Molecular Biology Module, Faculty of Medicine, Muhammadiyah University of Palembang for the academic year 2013/2014 to 2015/2016 showed that 45.2% of the questions had an ideal level based on DIFI. and 23.1% of the items had 100% DE value [6]. This study did not conduct a bivariate analysis of the relationship between the DE with DIFI and DI. Researchers have not found any publication of the item analysis that has been carried out at the Faculty of Medicine, State Islamic University Syarif Hidayatullah Jakarta, including the Sub Division of Orthopaedics and Traumatology. Researchers also find it difficult to obtain similar publications in other medical faculties, even though this is very important. If the multiple choice questions are not qualified, then the questions cannot measure the student's knowledge validly and reliably. Based on the elaboration of previous research, researchers are interested in conducting research on item analysis of the Musculoskeletal Module Sub Division of Orthopaedics and Traumatology, Faculty of Medicine, State Islamic University Syarif Hidayatullah Jakarta.

II. METHODS

Time and place

The observational analytic study with cross sectional design is held from March to October 2020 at Syarif Hidayatullah State Islamic University Faculty of Medicine, Jakarta.

Participants

The research data used secondary data obtained from the Medical Education Unit (MEU). The research samples are determined using the total sampling method in which there were 155 items with 620 distractors. The inclusion criteria were MCQs of the summative exam in Musculoskeletal Module at Orthopaedics and Traumatology Division from 2018 to 2019**.** The exclusion criteria were annuled item caused by wrong key option.

Statistical analysis

The data analysis was the univariate test to determine DIFI, DI, DE, and NFD from the summative exam and bivariate analysis to determine association between DE and DIFI, also DE and DI. Data were analyzed using IBM SPSS Statistics version 22.0.

III. RESULTS AND DISCUSSION

Univariat

1. Difficulty Item

Table 1. Distribution of Question Difficulty Levels in Summative Examinations

No.	Year	Summative Exam	Difficult	Acceptable	Easy	Total
1.	**2018**	**1**	3 (6%)	26 (52%)	21 (42%)	50 (100%)
2.	**2018**	**2**	4 (16%)	5 (20%)	16 (64%)	25 (100%)
3.	**2019**	**1**	3 (8,6%)	9 (25,7%)	23 (65,7%)	35 (100%)
4.	**2019**	**2**	2 (4,5%)	12 (26,5%)	31 (69%)	45 (100%)
Total	12 (7,75%)	52 (33,55%)	91 (58,7%)	155 (100%)		

Based on the results from the Table 1, the first summative exam in 2018 consists of 3 (6%) difficult, 26 (52%) acceptable, and 21 (42%) easy. Then in the second summative exam in 2018 consisting of 4 (16%) difficult, 5 (20%) acceptable, and 16 (64%) easy. Then in the first exam in 2019, it is divided into 3 (8.6%) difficult, 9 (25.7%) acceptable, and 23 (65.7%) easy. On the other hand, the second summative exam in 2019 has 2 (4.5%) difficult, 12 (26.5%) acceptable, and 31 (69%) easy. The results obtained from the summative exam based on DIFI consisted of 12 (7.75%) items that were categorized as difficult questions, 52 (33.55%) items were categorized as moderate questions, and 91 (58.7%) items were categorized as easy questions.

2. Discrimination Index

Table 2. Distribution of Discrimination Index in Summative Examinations

No.	Year	Summative Exam	Good	Poor	Total
1.	**2018**	**1**	30 (60%)	20 (40%)	50 (100%)
2.		**2**	10 (40%)	15 (60%)	25 (100%)
3.	**2019**	**1**	26 (74,3%)	9 (25,7%)	35 (100%)
4.		**2**	23 (51,1%)	22 (48,9%)	45 (100%)
Total	89 (57,4%)	66 (42,6%)	155		

								(100%)

Based on the Table 2, the first summative exam in 2018 consists of 30 (60%) good questions and 20 (40%) poor questions. Then in the second summative exam in 2018 consisted of 10 (40%) good and 15 (60%) poor. Then in the first summative exam in 2019 it was divided into 26 (74.3%) good and 9 (25.7%) poor, and in the second summative exam in 2019 consisting of 23 (51.1%) good and 22 (48.9%) are poor. From the 2018-2019 summative exam, it was found that the question categories were based on IDB parameters consisting of 89 items (57.4%) good and 66 items (42.6%) poor.

3. Non Functional Distractors

Table 3. Distribution of Non-Functional Distractors in Summative Examinations

No.	Year	Summative exam	NFD					Total
			0	1	2	3	4	
1.	**2018**	1	3 (6%)	12 (24%)	18 (36%)	11 (22%)	6 (12%)	50 (100%)
2.		2	3 (12%)	1 (4%)	8 (32%)	3 (12%)	10 (40%)	25 (100%)
3.	**2019**	1	1 (2,9%)	8 (22,9%)	7 (20%)	13 (37,1%)	6 (17,1%)	35 (100%)
4.		2	1 (2,2%)	4 (8,9%)	15 (33,3%)	17 (37,8%)	8 (17,8%)	45 (100%)
Total			8 (5,2%)	25 (16,1%)	48 (31%)	44 (28,4%)	30 (19,3%)	155 (100%)

The results of Table 3 based on NFD in 2018-2019, 8 (5.2%) items had 0 NFD. Then, the items with 1 NFD were 25 (16.1%) items. Then, 48 (31%) items had the number of 2 NFD. While the items with the number of 3 NFD were 44 (28.4%), and 30 (19.3%) items had the number of 4 NFD. The number of distractors on the summative exam is 620, consisting of 200 in the first summative exam in 2018, 100 in the second summative exam in 2018, 140 in the first summative exam in 2019, and 180 in the second summative in 2019. On the exam, there were also a number of functioning distractors (0 NFD) and distractors that are non-functioning (1 to 4 NFD). The first summative exam in 2018 had 95 (47.5%) functional distractors and 105 (52.5%) distractors that are non-functioning. The second summative exam in 2018 had 34 (34%) functioning distractors and 66 (66%) malfunctioning distractors. Then, there were 55 (39.3%) functioning distractors and 85 (60.7%) distractors did not function on the first summative exam item in 2019. Meanwhile, the second summative item in 2019 contained 63 (35%) functioning distractors and 117 (65%) malfunctioning distractors.

Thus, the total number of distractors on the summative exam questions that functioned was 247 (39.8%) and distractors that are non-functioning were 373 (60.2%). The summative exam has 8 (5.2%) high category items with 0 NFD. Then, 73 (47.1%) moderate category items that have NFD 1 to 2, and 74 (47.7%) questions low category with 3 and 4 NFD.

4. Distractor Efficiency

Table 4. Distribution of Distractor Efficiency in Summative Examinations

No.	Year	Summative Exam	DE

			High	Moderate	Low	Total
1.	**2018**	**1**	3 (6%)	30 (60%)	17 (34%)	50 (100%)
2.		**2**	3 (12%)	9 (36%)	13 (52%)	25 (100%)
3.	**2019**	**1**	1 (2,9%)	15 (42,9%)	19 (54,3%)	35 (100%)
4.		**2**	1 (2,2%)	19 (42,2%)	25 (55,6%)	45 (100%)
Total			8 (5,2%)	73 (47,1%)	74 (47,7%)	155 (100%)

Results based on DE obtained in the first summative exam in 2018 consisted of 3 (6%) high categories, 30 (60%) moderate categories, and 17 (34%) easy categories. Then, in the second summative exam in 2018 it has 3 (12%) high categories, 9 (36%) moderate categories, and 13 (52%) easy categories. Then, the first summative exam in 2019 is divided into 1 (2.9%) high category, 15 (42.9%) moderate category, and 19 (54.3%) easy category, and the second summative exam in 2019 consists of 1 (2.2%) high category, 19 (42.2%) moderate category, and 25 (55.6%) easy category. In the summative exam based on DE, there are 8 items (5.2%) in high category, 73 (47.1%) moderate category, and 74 (47.7%) low category.

Bivariat

1. Association between Distractor Efficiency and Difficulty Index

Table 5. Association between Distractor Efficiency and Difficulty Index

DE	DIFI				
	Difficult	**Acceptable**	**Easy**	**Total**	***P value***
	N	**N**	**N**	**N**	
Low	0 (0%)	8 (10,8%)	66 (89,2%)	74 (100%)	0.000
Moderate	8 (11%)	40 (54,8%)	25 (34,2%)	73 (100%)	
High	4 (50%)	4 (50%)	0 (0%)	8 (100%)	
Total	12 (7,75%)	52 (33,55%)	91 (58,7%)	155 (100%)	

There is a significant association ($P < 0.05$) between the two values. Based on the Table 5 results, it was found that in 74 summative exam items with low DE, 0 (0%) was difficult, 8 (11%) acceptable, and 66 (89.2%) easy based on DIFI. Then, from 73 summative exam items with moderate of DE, 8 (11%) were difficult, 40 (54.8%) acceptable, and 25 (34.2%) easy. Meanwhile, 8 items with high DE consisted of 4 (50%) difficult, 4 (50%) acceptable, and 0 (0%) easy.

2. Association between Distractor Efficiency and Discrimination Index

Table 6. Association between Distractor Efficiency and Discrimination Index

DE	DI			***P value***
	Good	**Poor**	**Total**	
	N	**N**	**N**	
Low	33 (44.6%)	41 (55.4%)	74 (100%)	0.008
Moderate	50 (68,5%)	23 (31,5%)	73 (100%)	
High	6 (75%)	2 (25%)	8 (100%)	
Total	89 (57.4%)	66 (42.6%)	155 (100%)	

Based on the results of the Table 6, it was found that the value of $P = 0.008$,

which means that there is a significant association ($P < 0.05$) between the two values. The results of the analysis of the association between DE and DI, found that from 74 summative exam items with low DE, 33 (44.6%) were good, and 41 (55.4%) were poor based on DI. Then, from 73 items with moderate DE, 50 (68.5%) were good, and 23 (31.5%) were poor. Then, on 8 items that had high DE category, it consisted of 6 (75%) good items, and 2 (25%) poor items.

IV. DISCUSSION

Difficulty Index

The difficulty index (DIFI) analysis of summative examinations in the Musculoskeletal Module at the Orthopaedics and Traumatology Subdivision (2018-2019) revealed a suboptimal distribution pattern compared to existing literature. Purwoko and Mundijo (2018) reported a more balanced distribution in the Molecular Biology and Genetic Module at Faculty of Medicine, Muhammadiyah University of Palembang, with 31% easy items, 45.2% acceptable items, and 23.8% difficult items among 299 questions analyzed [6]. Similarly, Rao (2016) found 85% acceptable items, 5% easy items, and 10% difficult items in 40 MCQs at the Department of Pathology, KS Hegde Medical Academy, India [7].

The present study demonstrated poor distribution of item difficulty, with 58.7% easy items, 33.55% acceptable items, and only 7.75% difficult items (n=12). This distribution fails to meet established standards for proportional difficulty spread. Arifin from the Department of Curriculum and Educational Technology, University of Education Indonesia, outlined optimal distributions as: (a) 25% difficult, 50% acceptable, 25% easy; (b) 20% difficult, 60% acceptable, 20% easy; or (c) 15% difficult, 70% acceptable, 15% easy [8].

This skewed distribution suggests inadequate consideration of item difficulty during examination construction. Supporting this interpretation, Sim and Rasiah found 53% acceptable questions among 50 MCQs at the Faculty of Medicine, University of Malaya, Malaysia, while Abdulghani reported that King Saud University, Riyadh, utilized predominantly acceptable items in both pre-test (62.5%) and post-test (76.3%) examinations [9]. The predominance of easy items may reflect the cognitive level distribution of medical curricula. Adams identified six cognitive levels: knowledge (remember), comprehension (understand), application (apply), analysis (analyze), evaluation (evaluate), and synthesis (create) [10] [11]. Multiple studies have documented that medical faculty curricula and learning objectives emphasize lower Bloom's taxonomy levels rather than higher-order thinking skills. Consequently, a higher proportion of easy items based on DIFI may be expected [10]. However, when students perform poorly on examinations dominated by recall-level questions, alternative explanations must be considered, including inadequate student preparation, insufficient understanding of tested material, misinterpretation due to technical item flaws, or poor student selection. A more proportional distribution of cognitive levels facilitates deeper learning and enhances knowledge and skill transfer across diverse tasks and contexts [10].

Difficult items in this study exhibited DIFI values ranging from 0.095 to 0.295. Notably, no items achieved DIFI = 0.00, indicating that at least some students answered every question correctly. Acceptable items ranged from 0.326 to 0.695 and can be retained for future use. Easy items ranged from 0.705 to 1.00, with DIFI = 1.00 indicating extremely easy items answered correctly by all examinees. Items with DIFI

values of 0 or 1 represent extreme difficulty or ease, respectively.

Multiple factors contribute to extreme item difficulty. Tarrant et al. (2008) identified common item construction errors that render questions excessively difficult or easy, categorizing technical flaws into testwiseness and irrelevant difficulty [12]. Testwiseness factors that facilitate correct responses include grammatical cues, logical cues, word repetition, longest option being correct, absolute terms, implausible distractors, convergence strategies, and "all of the above" or "none of the above" options [12]. Conversely, items may become excessively difficult when they exceed taught learning outcomes [10]. Student cohort quality also influences item difficulty categorization. Irrelevant difficulty factors include vague terminology, poor sequencing of numeric data, unnecessary information, unclear or ambiguous content, multiple or no correct answers, failure of the "hand cover test," and negative stems (not, except).

Extremely difficult or easy items compromise validity and reliability of summative examinations, meaning student results may not accurately represent true competence [13]. Items categorized as difficult or easy should be evaluated proportionally before reuse. Items with technical defects should be revised before storage in item banks. Extreme items (DIFI = 0 or 1) should be eliminated.

Discrimination Index

Discrimination index (DI) analysis results align with Rao's findings of 60% very good items (D > 0.4), 10% good items (D = 0.3-0.39), and 15% acceptable items (D = 0.2-0.29), totaling 85% good items (D > 0.2) and 15% poor items (D < 0.2).[7] Kheyami et al. (2018) reported 68% good questions and 32% poor questions among 800 items at the Department of Pediatrics, Arabian Gulf University, Bahrain [14].

Items with good DI ranged from 0.217 to 0.652. Higher DI values indicate superior discrimination between the upper 27.5% and lower 27.5% of examinees. Poor DI items ranged from -0.192 to 0.192, subdivided into those amenable to revision and those requiring elimination. Negative DI values indicate more lower-group participants answered correctly than upper-group participants. A DI value of zero indicates equal correct responses between groups; four items exhibited D = 0 in this study. Items with D = 0 typically represent extremely easy questions, often due to lower-order thinking requirements [12]. Testwiseness factors also enable correct responses without adequate competence. However, D = 0 may also occur when all students have achieved predetermined competencies. Serial examinations can distinguish whether full competency achievement or item defects explain this finding [15].

Negative DI scores suggest reversed examinee quality discrimination. Poorly structured, ambiguous items. For example, when stems, questions, and answer options lack coherence—may mislead higher-performing students toward incorrect options, while random guessing by lower-performing students may contribute to negative DIFI scores [16] [17].

Items in the good category should be stored in item banks for future examinations. Poor DI items that remain improvable (0 < D < 0.2) should be revised. Items with D = 0 or negative values should be eliminated due to their impact on test validity and reliability in differentiating student competence, particularly between upper and lower performance groups [15].

Non-Functional Distractors

The NFD analysis differed from Kheyami et al. (2018), who reported 48.4% high NFD category (NFD = 0), 46.7% moderate category (NFD = 1-2), and 5% low

category (NFD = 3-4) among 800 MCQs [14]. Sajjad (2020) found predominantly moderate-category questions (42%, n=75) among 181 MCQs in surgical class examinations at Riphah International University, Pakistan. The present study identified 5.2% (n=8) high-category NFD items, 47.1% (n=73) moderate-category items, and 47.7% (n=74) low-category items. The paucity of NFD = 0 items reflects poor distractor quality or defects. Item analysis revealed testwiseness and irrelevant difficulty factors that enabled students to easily identify incorrect options [12].

Comparison of studies using different distractor numbers provides additional insight. Burud (2019) at International Medical University, Malaysia, employed three distractors, while Purwoko and Mundijo (2018) at Muhammadiyah Palembang University utilized four distractors. Both reported higher proportions of items with NFD = 1. Kheyami et al. (2018) from Arabian Gulf University, Bahrain, using four distractors, found the largest proportion was items with NFD = 0 [6] [9] [14].

Optimal items have NFD = 0 and warrant storage in item banks. Items with NFD = 1-4 require revision before reuse. Providing feedback to item writers through MCQ analysis helps identify NFD errors. Based on this analysis, 5.16% (n=8) items can be stored, while 94.84% (n=147) require revision [18].

Distractor Efficiency

The DE analysis differed from Sajjad (2020), who found 42% (n=75) moderate DE items, 38% (n=69) high DE items, and 20% (n=37) low DE items among 181 MCQs in surgical class examinations [19].

Low DE items typically exhibit easy DIFI scores ($p > 0.7$) or poor DI scores ($D < 0.2$) [24]. Poor distractor quality, including testwiseness and irrelevant difficulty factors, causes low DE by enabling examinees to easily identify incorrect options [12]. High DE items should be stored in item banks. Moderate and low DE items require revision, with low DE items first revised to moderate level. Distractor revision involves replacing ineffective options with more relevant, plausible alternatives. Abdalla (2011) recommended establishing consensus on answer keys before writing plausible distractors directly related to item stems (based on clinical symptoms, diagnoses, etc.). Panel review of MCQs before administration is preferable [20].

Bivariate Analysis

1. Association Between Distractor Efficiency and Difficulty Index

Analysis of meaningful relationships revealed 89.2% (n=66) of low DE items were easy items among 74 MCQs in this category. Additionally, 54.8% (n=40) of moderate DE items were acceptable items among 73 MCQs, while 50% (n=4) of high DE items were difficult items among 8 MCQs. Lower DE values correlated with easier DIFI classifications, and vice versa.

These findings align with Kheyami et al. (2018), who demonstrated significant association between DE and DIFI ($p < 0.050$) in the Department of Pediatrics at Arabian Gulf University, Bahrain, concluding that more difficult MCQs have greater DE values [14]. Abdulghani et al. (2016) from King Saud University found significant relationships between NFD number and DI in Family Medicine examinations, with more NFDs correlating with easier items or higher DIFI values [18].

2. Between Distractor Efficiency and Discrimination Index

Analysis revealed 56.8% (n=42) of low DE items had poor DI among 74 MCQs. Furthermore, 68.5% (n=50) of moderate DE items had good DI among 73 MCQs, while 75% (n=6) of high DE items had good DI among 8 MCQs. Higher DE

values correlated with better DI values, and vice versa. Good items with moderate DE require distractor revision to achieve higher quality for future examinations.

These results support Kheyami et al. (2018), who demonstrated significant association between DE and DI ($p < 0.050$), concluding that items with good DI have greater DE values than items with poor DI [21].

V. CONCLUSION

This study analyzed 155 multiple-choice questions with 620 distractors from the Musculoskeletal Module at Syarif Hidayatullah State Islamic University Faculty of Medicine, Jakarta (2018-2019). The difficulty index analysis revealed that 12 items (7.75%) were difficult, 52 items (33.55%) were moderate, and 91 items (58.7%) were easy. The discrimination index showed that 89 items (57.4%) had good discriminatory power, while 66 items (42.6%) showed poor discrimination. Regarding distractor efficiency, 82 items (52.9%) demonstrated high efficiency, whereas 73 items (47.1%) showed low efficiency, with 573 distractors (60.2%) classified as non-functional.

Statistical analysis revealed that distractor efficiency was significantly associated with both difficulty index and discrimination index ($p <0.05$). These findings indicate that while the assessment items demonstrated adequate difficulty levels and reasonable discriminatory power, there is considerable room for improvement in distractor functionality. Future item development should prioritize constructing functional distractors to optimize the overall quality and validity of assessments in the Musculoskeletal Module.

Conflicts of Interest

The authors declare that there are no conflicts of interest related to the preparation, content, or publication of this document.

Funding Sources

This work did not receive any specific fund.

Aknowledgements

The authors would like to express their sincere appreciation to the Medical Education Unit, Faculty of Medicine UIN Jakarta, for their valuable support and for providing access to the data used in this work.

REFERENCES

[1] Gronlund NE, Waugh CK. Measurement and Evaluation in Teaching. New York: McMillan Publishing; 2009. h. 1–10.

[2] Amin Z, Seng CY, Eng KH. Practical Guide to Medical Student Assessment. Singapore: World Scientific Publishing; 2006. h. 22–24.

[3] Lynch DC, Swing ST. Key Considerations for Selecting Assessment Instruments and Implementing Assessment Systems. Accreditation Council of Graduate Medical Education (ACGME). Available on: http://www.acgme.org/outcome/assess/keyConsider.asp

[4] Shumway JM, Harden RM. AMEE Guide No. 25: The Assessment of Learning Outcome for the Competent and Reflective Physician. Medical Teacher. 2003;25(6):578–581.

[5] Mukherjee P, Lahiri SK. Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an Assessment in a Medical College of Kolkata, West Bengal. IOSR ***Journal of Dental and Medical Sciences***. 2015;14(12):47–52.

[6] Purwoko M, Mundijo M, T T. Evaluating the Use of MCQ as an Assessment Method in a Medical School for Assessing Medical Students in the Competence-Based Curriculum. ***The Indonesian Journal of Medical Education***. 2018;7(1):44–48.

[7] Rao C, Prasad K, Sajitha K, Harish P, Shetty J. Item Analysis of Multiple Choice Questions: Assessing an Assessment Tool in Medical Students. ***International Journal of Educational and Psychological Researches***. 2016;2(4):201–204. Available on: http://www.ijeprjournal.org

[8] Arifin Z. Evaluasi Pembelajaran. Bandung: PT Remaja Rosdakarya; 2017. h. 2, 266–283.

[9] Burud I, Nagandla K, Agarwal P. Impact of Distractors in Item Analysis of Multiple Choice Questions. ***International Journal of Research in Medical Sciences***. 2019;7(4):1136–1139.

[10] Adams N. Bloom's Taxonomy of Cognitive Learning Objectives. ***Journal of the Medical Library Association***. 2015;103(3):152–153.

[11] Anderson LW, Krathwohl DR, et al. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Boston, MA: Pearson Education Group; 2013.

[12] Abdalla ME, Gaffar AM, Sulaiman RA. Constructing A-Type Multiple Choice Questions (MCQs): Step by Step Manual. United Arab Emirates; 2011. h. 15–22.

[13] Arifin Z. Evaluasi Pembelajaran. Bandung: PT Remaja Rosdakarya; 2017. h. 2, 266–283.

[14] Kheyami D, Jaradat A, Al-Syihbani T, Ali FA. Item Analysis of Multiple Choice Questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain. Sultan Qaboos University ***Medical Journal***. 2018;18(1):67–74. Available on: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5892816/

[15] Al-Wardy NM. Assessment Methods in Undergraduate Medical Education. Sultan Qaboos University ***Medical Journal***. 2010;10(2). Available on: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074721/

[16] Panitia Nasional UKMPPD. Buku Pedoman Pelaksanaan UKMPPD. Jakarta: Panitia Nasional UKMPPD; 2014. h. 4–5.

[17] Surapranata S. Analisis, Validitas, Reliabilitas, dan Interpretasi Hasil Ujian. Bandung: PT Remaja Rosdakarya; 2006. h. 10.

[18] Abdulghani HM, Ahmad F, Ponnamperuma GG, Khalil MS, Aldees A. The Relationship Between Non-Functioning Distractors and Item Difficulty of Multiple Choice Questions: A Descriptive Analysis. ***Journal of Health Specialties***. 2014;2(2):148–151.

[19] Sajjad M, Iltaf S, Khan RA. Nonfunctional Distractor Analysis: An Indicator for Quality of Multiple Choice Questions. ***Pakistan Journal of Medical Sciences***. 2020;36(5):983–986.

[20] Health Professional Education Quality Project Direktorat Jenderal Pendidikan Tinggi Kementerian Pendidikan Nasional. Pedoman Administrasi Bank Soal Uji Kompetensi Nasional Tenaga Kesehatan Indonesia. Jakarta: Kementerian Pendidikan; 2010. h. 3–21.

[21] Menon AR, Kannambra PN. Item Analysis to Identify Quality Multiple Choice Questions. ***National Journal of Laboratory Medicine***. 2017;6(2):7–10.