JURNAL PENDIDIKAN DAN PENGAJARAN
Volume 57 Nomor 1 2024, 47-56
E-ISSN: 2549-2608; P-ISSN: 2301-7821
DOI: https://doi.org/10.23887/jpp.v57i1.63314

Multiple-Choice Questions in Basic Biomedical Science Module

Made Bayu Permasutha[1*], Gandes Retno Rahayu[3], Made Kurnia Widiastuti Giri[4], I Dewa Agung Gde Fanji Pradiptha[5]

[1,4,5] Universitas Pendidikan Ganesha, Indonesia
[2] FAIMER Regional Institute of Indonesia for Educational Development and Leadership (FRIENDSHIP) Fellowship, Indonesia
[3] Departement of Medical Education, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Indonesia
*Corresponding author: bayu.permasutha@undiksha.ac.id

Abstrak

Proses evaluasi pendidikan kedokteran melibatkan evaluasi pengetahuan, keterampilan, dan sikap yang sesuai dengan capaian dan kompetensi yang ingin dicapai. Instrumen multiple choice questions (MCQs) merupakan instrument penilaian yang sering digunakan di ranah kedokteran. Instrumen MCQs hingga saat ini juga diimplementasikan dalam ujian kompetensi dokter nasional. Sehingga menjaga kualitas MCQs di level fakultas sangat penting dalam menjaga kualitas lulusan dokter. Dalam penelitian ini dilakukan evaluasi terhadap 250 MCQs di modul biomedik dasar yang kemudian dilakukan beberapa analisis terhadap beberapa kriteria. Dari hasil analisis didapatkan nilai Kr-20 sebesar >0,8 pada ketiga modul. Item Difficulty Index (DIF-I) pada kategori ideal didapatkan 33 (36,7%), 29 (38,7%), dan 34 (39,5%) secara berurutan. Hasil kajian terhadap Item Discrimination Index (DI) dalam kategori ideal didapatkan sebesar 63,3%, 77,3% dan 69,4% secara berurutan. Hasil studi ini memberikan gambaran bahwa masih terdapat MCQs yang tidak ideal dan perlu mendapat perhatian untuk perbaikan ke depannya. Evaluasi secara berkala dan pelatihan pembuatan komponen soal pilihan ganda yang terstandar perlu untuk direncanakan di lingkup fakultas.

Kata Kunci: *Pertanyaan Pilihan Ganda, Ilmu Biomedis Dasar, Indeks Diskriminasi Item.*

Abstract

The evaluation process in medical education involves evaluating knowledge, skills, and attitudes based on the achievements and competencies to be achieved. The multiple-choice question (MCQ) instrument is an assessment instrument often used in the medical field. So far, the MCQs have been implemented in Indonesia's national medical competency examination. Furthermore, maintaining the quality of MCQs at the faculty level is very important to maintain the quality of medical graduates. In this study, an evaluation was carried out on 250 items of MCQs in three basic biomedical modules, followed by analyses of the MCQs characteristics, item difficulty index (DIF-I), and item discrimination index (DI). The analysis found that the Kr-20 value was >0.8 in the three modules. Analysis of the item difficulty index (DIF-I) in the ideal category obtained 33 (36.7%), 29 (38.7%), and 34 (39.5%), respectively. The ideal category's item discrimination index (DI) was 63.3%, 77.3%, and 69.4%, respectively. The results of this study illustrate that there are still MCQs that are not ideal and need attention for future improvements. These results prove that more work must be done to improve the standard of MCQs used in medical examinations. Periodic evaluation and training on making standardized multiple-choice question components need to be planned within the faculty.

Keywords: Multiple Choice Questions, Basic Biomedical Sciences, Item Discrimination Index.

History:
Received : June 11, 2023
Accepted : December 20, 2023
Published : April 25, 2024

Publisher: Undiksha Press

1. INTRODUCTION

The purpose of the learning evaluation process in medical education is to assess the attainment of learning objectives, comprehension of the learning process, and standard competencies, as well as to develop and evaluate educational programs (Boulet & Durning, 2019; Tavakol & Dennick, 2017). In recent decades, medical institutions have prioritized conducting competency evaluations that are accurate, dependable, and time-efficient. The evaluation process has three primary goals: to optimize students' abilities through motivation and direction for future learning; to safeguard society through the competency assessment process for medical graduates; and to provide students with a foundation for pursuing advanced degrees (Lockyer et al., 2017; Ten Cate & Regehr, 2019). One common issue encountered at the faculty level pertains to the absence of exam item evaluation. This phenomenon may result in a lack of training for medical students to effectively address highly complicated questions in alignment with the requirements of the Indonesian Medical Doctors National Competency Examination (IMDNCE) in the future. In a general sense, this study assesses the efficacy of multiple-choice questions that have undergone rigorous examination with students using various analyses.

Medical education involves the development of knowledge, abilities, and attitudes (Harden, 2018; Rodríguez, 2014). Consequently, the instruments used to evaluate these components will vary. Previous research identified several evaluation instruments used to evaluate 'knows' and 'knows how," "shows how," and 'does' (Boulet & Durning, 2019; Tavakol & Dennick, 2017). Multiple-choice questions (MCQs) are one of the instruments used to evaluate 'knows' and 'knows how.' Due to their high validity and reliability, ease of assessment, and efficiency in mass-assessing candidates, MCQs are increasingly used in medical institutions (Hijji, 2017; Kumar et al., 2021). In medical schools, providing students with adequate and accurate information and enhancing their basic knowledge and practical abilities is crucial. The assessment conducted during teaching and learning practice serves multiple purposes. Not only does this ensure that students are able to comprehend the material, but it also allows us to evaluate how effective our teaching strategies are. Therefore, assessment procedures should be reliable and efficient (Boulet & Durning, 2019; Tavakol & Dennick, 2017). Continuous analyses of student assessment methodologies should be a crucial step for enhancing the knowledge of the students and the quality of examinations. Previously defined pre- and post-validation assessment methods exist for analyzing the queries formulated. Prior to conducting the assessment, the experts should evaluate the relevancy of the topics and the suitability of the MCQ structure, including stem and options, as part of the pre-validation process. The post-validation procedure is essentially a statistical technique known as item analysis. This is a valuable, relatively uncomplicated, and effective method for determining the reliability and validity of multiple-choice questions (Kumar et al., 2021; Kurtz et al., 2019). This is beneficial in three ways. The difficulty index (DIF-I) indicates whether the MCQs provided to a student are challenging or simple to answer. Second, it can distinguish between students who have excellent subject knowledge and those with poor performance. The term for this is the discrimination index (DI). Thirdly, it assists the subject expert in evaluating the validity of incorrect options (distractors). The term for this is distractor efficiency (DE). Overall, this analysis provides the evaluator with guidelines for improving the MCQs before the next examination.

The Indonesian Medical Doctors National Competency Examination, also known as the *Uji Kompetensi Mahasiswa Program Profesi Dokter* (UKMPPD), is the final evaluation for future doctors to maintain the quality of graduates in the medical profession. It is also a form of protection for the community when utilizing medical services in Indonesia. Evaluation of the UKMPPD has comprised of two stages: the computer-based test (CBT) with the MCQ instrument and clinical competency evaluation with the Objective Structured Clinical Examination (OSCE) instrument (Darmayani, 2022; Utomo, P. S. et al., 2022).

A well-constructed MCQ instrument is taxonomically augmented to assess higher-order cognitive abilities, such as applying knowledge, interpretation, and synthesis, rather than just recalling isolated facts. Due to the significance of sustaining the quality of multiple-choice questions, various indicators can be used to evaluate the quality of MCQ items. Cognitive levels, item writing flaws (IWFs), the item difficulty index (DIF-I), and the item discrimination index (DI) are examples of these indicators (Baig et al., 2014; Christian et al., 2017; Xu et al., 2016). Item analysis provides a method for evaluating the content of questions by determining how relevant they were to respondents and how accurately they measured their ability. Item analysis provides objective evidence of student progress toward the subject's concepts and simplifies the topic for students (Biswas et al., 2015; León et al., 2023). Frequently, it is challenging for the lecturers to evaluate the quality of items repeatedly used to evaluate students' performance. Item analysis also provides a method for repeatedly reusing items in various instruments with prior knowledge of their performance. In classical analysis, the difficulty of a single-response multiple-choice question is merely the proportion of incorrect responses. The item consists of a stem (question) and five alternatives, of which one is the correct response and the others are incorrect.

The novelty of the study is a thorough evaluation of the effectiveness of multiple-choice questions (MCQs) that have undergone rigorous examination with students using various analyses. The main objective is to ensure that the evaluation instrument meets the quality standards needed to measure students' ability to master the material and competence of medical standards. By taking into account these criteria, this research makes an important contribution in improving the quality of examinations and student evaluations in medical education. The Faculty of Medicine at Universitas Pendidikan Ganesha was established in 2018 and implemented an MCQ-based evaluation instrument for each academic module. However, assessment instruments must be valid, reliable, and capable of measuring the various facets of professional competency. This study was conducted to evaluate MCQ items in Basic Biomedical Sciences modules by identifying the characteristics of MCQs as well as the item difficulty index and item discrimination index.

2. METHODS

This research assessed a collection of MCQs used in the final exams for the basic biomedical 1, 2, and 3 modules in the undergraduate medical level's first semester of the 2020–2021 academic year. Sampling was conducted on all MCQs in the 2020–2021 academic year. The MCQs used were question types with 4-5 answer items with the single best answer (SBA)-summative type MCQs. Two hundred fifty questions from three modules (90 questions in basic biomedical module 1, 75 in basic biomedical module 2, and 85 in basic biomedical module 3) were analyzed according to the research objectives. Furthermore, this research did not directly involve human subjects. This research has received ethical clearance from ethics committee Faculty of Medicine Universitas Pendidikan Ganesha with reference number 017/UN48.24.11/LT/2023.

MCQ item analysis was performed using IBM SPSS Statistics software version 26. Correct answer responses were coded as 1, and incorrect answers were coded as 0. DIF-I, or p-value (proportion or percentage value), was used to correctly calculate the percentage of students who answered MCQ items. DIF-I has a score range of 0.0 to 1.0, with a score of 0.0 indicating a question that is too easy and a value of 1.0 indicating a question that is too difficult. The ideal DIF-I range is 0.3–0.7. The interpretations of the item difficulty index values are shown in Table 1. Moreover, the item difficulty index formula is described as follows:

$$\frac{number\ of\ students\ answering\ item\ correctly}{total\ number\ of\ students} \qquad (1)$$

Table 1. Interpretation of Item Difficulty Index

Interpretation of Items	DIF-I Values
Too easy	>0,7
Average	0,3-0,7
Too difficult	<0,3

The Item Discrimination Index (DI) calculation to determine whether an item question can distinguish between students who have studied well and those who have not. The DI has a range of values from -1 to 1, where a positive value of one indicates that the item can distinguish between students who study well and those who do not. A value of zero means that the item is unable to distinguish performance between these two groups. Meanwhile, a negative value of 1 indicates that the item can discriminate well but is reversed, where students with good performance will answer incorrectly while students who do not study will answer correctly. The interpretation of DI values is described in Table 2. In the calculation, 27% of students with the highest scores will be identified as the upper group and 27% of students with the lowest scores as the lower group. The formula for the item discrimination index is described as follows.

$$\frac{upper\ group - lower\ group}{27\%\ of\ total\ students} \qquad (2)$$

Table 2. Interpretation of Item Discrimination Index

Interpretation of Items	DI Values
Excellent Item	0,4-1
Good Item	0,25-0,39
Marginal Item	0-0,24
Bad Item	<0

3. RESULTS AND DISCUSSION

Result

The characteristic analysis of 250 MCQs is shown in Table 3. The characteristics of the three modules completed by 56 students showed a reliability value (Kuder-Richardson 20) of 0.98, 0.91, and 0.21, respectively. However, two items in basic biomedical modules 1 and 3 showed zero variance and were excluded from the characteristic analysis.

Table 3. Characteristics of MCQs

Item Characteristics	Basic Biomedical Module 1	Basic Biomedical Module 2	Basic Biomedical Module 3
Number of items	90	75	85
Number of examines	56	56	56
Mean test score	59.64±12.45	60.38±15.39	63.26±15.28
Range of test score	35.56-83.33	28.00-84.00	31.76-85.88
Kuder-Richardson 20 (Kr-20)	0.89	0.91	0.92*

The DIF-I analysis of MCQs on each module is shown in Table 4, Table 5, and Table 6. The DIF-I analysis showed that less than 40% of the MCQs item in each biomedical module indicated the ideal value of 0.3-0.7. These results showed that some items were either too easy or difficult to solve.

Table 4. Item Difficulty Index of MCQs

Difficulty index	Basic Biomedical Module 1	
(P value)	No. of items (n=90)	Mean P value
>0,7 (Too easy)	38 (42.2%)	85.43±9.09
0,3-0,7 (Average)	33 (36.7%)	54.11±12.02
<0,3 (Too difficult)	19 (21.1%)	17.67±7.14

Table 5. Item Difficulty Index of MCQs

Difficulty index (P value)	Basic Biomedical Module 2	
	No. of items (n=75)	**Mean P value**
>0.7 (Too easy)	33 (44%)	81.76±6.10
0.3-0.7 (Average)	29 (38.7%)	54.86±12.51
<0.3 (Too difficult)	13 (17.3%)	18.41±7.16

Table 6. Item Difficulty Index of MCQs

Difficulty index (P value)	Basic Biomedical Module 3	
	No. of items (n=85)	**Mean P value**
>0.7 (Too easy)	42 (48.8%)	82.65±8.21
0.3-0.7 (Average)	34 (39.5%)	52.28±13.16
<0.3 (Too difficult)	9 (10.5%)	14.29±11.01

Meanwhile, DI analysis varied for each biomedical module. For each module, items belonging to the ideal category (0.25-1) were 63.3%, 77.3%, and 69.4%, respectively. Negative DI values were also found in each module with a percentage of ≤8%, and the results are shown in Table 7.

Table 7. Item Discrimination Index of MCQs

Discrimination index	Basic Biomedical Module 1	Basic Biomedical Module 2	Basic Biomedical Module 3
	No. of items (n=90)	**No. of items (n=75)**	**No. of items (n=85)**
0.40-1 (Excellent)	28 (31.1%)	36 (48%)	43 (50.6%)
0.25-0.39 (Good)	29 (32.2%)	22 (29.3%)	16 (18.8%)
0-0.24 (Marginal)	26 (28.9%)	11 (14.7%)	20 (23.5%)
<0 (Poor)	7 (7.8%)	6 (8%)	6 (7.1%)

Discussions

Assessing student learning is an integral part of the educational process. Several evaluation models are used in education, including goal-oriented evaluation models, goal-free evaluation models, formative-summative evaluations, and countenance evaluation models (Darodjat & Wahyudhiana, 2015; Mardiah, M. & Syarifudin, 2018). In medical education, the formative-summative evaluation model is the most frequently used. The formative evaluation assesses the progress of the learning process and provides feedback to students, while the summative evaluation is used to evaluate the outcome of student learning. Furthermore, the summative evaluation helps to determine which students are qualified to continue to the next level or which students need to repeat the learning process (Kibble, 2017; Shafira, 2015). Several instruments can be used to evaluate medical students in formative or summative. Previous research categorized the evaluation instruments often used in the medical field as follows: Evaluation instruments to assess 'knows' and 'knows how' by using oral examinations/vivas, long essay questions (LEQ), short answer questions (SAQ), multiple choice questions (MCQs), extended matching items (EMI), and key features test (KF); Evaluation instruments to assess 'shows how' using long and short cases and objective structured clinical examination (OSCE); The instruments for assessing 'does' use a mini clinical evaluation exercise (Mini-CEX), direct observation of procedural skills (DOPS), clinical work sampling (CWS), checklist, 360-degree evaluation, logbook, and portfolio

(Boulet & Durning, 2019; Tavakol & Dennick, 2017).

Furthermore, well-constructed MCQs are the preferred choice in several medical faculties for formative and summative examinations. The examples of MCQs used in a formative examination are progress tests. The examples of MCQs used in the summative examination are final exams in each module, joint state tests, and national competency examinations (Mahda et al., 2023; Pugh & Regehr, 2016). With the appropriate instruction, training, and resources, multiple-choice questions (MCQs) can be used to assess students' higher thinking abilities (Donnelly, 2014; Zaidi et al., 2018). Comparing pre-training and post-training MCQ-based scores in intervention and non-intervention groups, Dellinges and Curtis discovered that a one-hour MCQ teaching workshop for 24 dental faculty increased the quality of in-house MCQs (AlFaris et al., 2015; Dellinges & Curtis, 2017). Previous study revealed that constructing more difficult MCQs involving problem-solving (level three) in clinical fields was more straightforward than in basic medical science modules and superior to other question types. In a study of 50 multiple-choice questions, similar research found that 60% of the questions focused on applying the knowledge plane, 28% on recalling information (level one), and only 6% on interpretation of data (level two) (Gupta et al., 2020; Przymuszała et al., 2020).

The analysis of MCQ characteristics in the three modules showed Kr-20 values of 0.89, 0.91, and 0.92, respectively. The Kr-20 value is the average of all correlations that aim to produce a reliability coefficient, indicating internal consistency and reliability at a range of 0 to 1 (Kaur et al., 2016; Rahma et al., 2017). A Kr-20 coefficient value of 0.8-1 is acceptable. Higher values highlight greater consistency and homogeneous reliability (e.g., values >0.9), and coefficients below 0.8 indicate that the entire test is unreliable and inconsistent. Therefore, based on the results, MCQs in the three modules showed good internal consistency and homogeneous reliability value. The ideal values for DIF-I and DI in MCQs are expected to be in the range of 0.3-0.7 for DIF-I and >0.25 for DI (Bhattacherjee et al., 2022; D’Sa & Visbal-Dionaldo, 2017; Kumar et al., 2021). In the DIF-I analysis for the three modules, the ideal results were 33 items (36.7%), 29 items (38.7%), and 34 items (39.5%), respectively. Furthermore, more than 60% of MCQs were non-ideal questions, with more than 40% comprising easy questions and less than 22% constituting difficult questions. Previous research stated that the DIF-I average tended to decrease in several questions with high cognitive levels, underscoring the importance of the lecturers in improving students' higher-order thinking skills (Daryono et al., 2020; Liew et al., 2021).

Furthermore, the DI results obtained the ideal category (0.25-1) of 63.3%, 77.3%, and 69.4%, respectively. The results also obtained a DI value below zero, indicating that the items needed to be adjusted to sufficiently distinguish between students who studied well and those who did not. Previous study observed a low DI tendency in MCQs with writing flaws in the stem, lead-in, and multiple-choice answers (Abdulghani et al., 2015; Rush et al., 2016). These results highlight the importance of adhering to writing standards for MCQs, particularly in content concerns, formatting concerns, style concerns, writing the stem, and writing the choices (Butler, 2018; Xu et al., 2016). Based on the preceding discourse, it is advisable to consider eliminating or modifying questions that exhibit a negative discrimination index and fall inside the easy group. The need for this change arises from the potential influence of simplistic and subpar questions on the accuracy of learning accomplishment assessments, hence impacting the outcomes of learning assessments (Adam et al., 2021; Nojomi & Mahmoudi, 2022). According to similar study, well-designed multiple-choice question (MCQ) assessments have the capacity to evaluate cognitive abilities at higher levels of Bloom's taxonomy, including data interpretation, data synthesis, and knowledge application, surpassing mere memory of instructional content (Adiga et al., 2021; Elgadal & Mariod, 2021). Several factors may restrict the generalizability of the results of this study. The study only evaluated the results of three modules. In addition, item writing flaws and distractor efficiency were unavailable.

4. CONCLUSION

The current study suggests a need to improve the quality of our assessment because some MCQ items still do not meet the ideal category. This will reduce the exam's validity and force students to adopt surface learning strategies, which are not conducive to lifelong learning. The utilization of inadequate questions for the test could negatively impact the overall validity of the assessment and compel students to employ superficial learning approaches, which are not favorable to the development of lifetime learning skills. It is essential to undertake the revision or elimination of inadequate questions in order to prevent their inclusion in future exams or storage within the question repository. Academics should practice MCQ item analysis more frequently because it provides insight into the questions' quality. This study's findings emphasized the importance of item analysis, including analysis of the difficulty and discrimination indices, which are frequently neglected in such examinations. Incorporating items with a moderate level of difficulty and a high level of discrimination into tests will enhance the queries' validity and effectiveness.

5. ACKNOWLEDGE

The author appreciates the coordinator of the medical study program, Faculty of Medicine Universitas Pendidikan Ganesha, for providing access to questions and answers bank in the three basic biomedical modules. The author is also grateful to all FAIMER fellows and supervisors who provided guidance in this research aimed at enhancing the capacity of medical education. Author also wants to thank Ganesha University of Education for the financial support (reference number 181/UN48.16/LT/2023) to complete this research. The funding agencies had no role in study design, data collection, data analysis, interpretation of data, report’s writing, and in submitting the article for publication.

6. REFERENCES

Abdulghani, H. M., Ahmad, F., Irshad, M., Khalil, M. S., Al-Shaikh, G. K., Syed, S., & Haque, S. (2015). Faculty development programs improve the quality of multiple choice questions items’ writing. *Scientific Reports, 5*(1). https://doi.org/10.1038/srep09556.

Adam, S. K., Idris, F., Kassim, P. S. J., Zakaria, N. F., & Hod, R. (2021). Multiple Choice Questions with Different Numbers of Options in University Putra Malaysia Undergraduate Medical Program: A Comparative Analysis in 2017 and 2018. *Journal of Medical Education*, *20*(2). https://doi.org/10.5812/jme.116834.

Adiga, M. N. S., Acharya, S., & Holla, R. (2021). Item analysis of multiple-choice questions in pharmacology in an Indian Medical School. *Journal of Health and Allied Sciences NU*, *11*(3), 130–135. https://doi.org/10.1055/s-0041-1722822.

AlFaris, E., Naeem, N., Irfan, F., Qureshi, R., Saad, H., Al Sadhan, R. E., & Van der Vleuten, C. (2015). A One-Day Dental Faculty Workshop in Writing Multiple-Choice Questions: An Impact Evaluation. *Journal of Dental Education*, *79*(11), 1305–1313. https://doi.org/10.1002/j.0022-0337.2015.79.11.tb06026.x.

Baig, M., Ali, S. K., Ali, S., & Huda, N. (2014). Quality evaluation of assessment tools: OSPE, SEQ & MCQ. *Pakistan Journal of Medical Sciences*, *30*(1), 3–6. https://doi.org/10.12669/pjms.301.4458.

Bhattacherjee, S., Mukherjee, A., Bhandari, K., & Rout, A. J. (2022). Evaluation of Multiple-Choice Questions by Item Analysis, from an Online Internal Assessment of 6(th) Semester Medical Students in a Rural Medical College, West Bengal. *Indian Journal of Community Medicine*, *47*(1), 92–95. https://doi.org/10.4103/ijcm.ijcm_1156_21

Biswas, S. S., Jain, V., Agrawal, V., & Bindra, M. (2015). Small group learning: effect on item analysis and accuracy of self-assessment of medical students. *Education for Health*, *28*(1), 16–21. https://doi.org/10.4103/1357-6283.161836.

Boulet, J. R., & Durning, S. J. (2019). What we measure… and what we should measure in medical education. *Medical Education*, *53*(1), 86–94. https://doi.org/10.1111/medu.13652.

Butler, A. C. (2018). Multiple-Choice Testing in Education: Are the Best Practices for Assessment Also Good for Learning? *Journal of Applied Research in Memory and Cognition*, *7*(3), 323–331. https://doi.org/10.1016/j.jarmac.2018.07.002.

Christian, D. S., Prajapati, A. C., Rana, B. M., & Dave, V. R. (2017). Evaluation of multiple choice questions using item analysis tool: a study from a medical institute of Ahmedabad, Gujarat. *International Journal Of Community Medicine And Public Health*, *4*(6), 1876–1881. https://doi.org/10.18203/2394-6040.ijcmph20172004.

D’Sa, J. L., & Visbal-Dionaldo, M. L. (2017). Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency. *International Journal of Nursing Education*, *9*(3). https://doi.org/10.5958/0974-9357.2017.00079.4.

Darmayani, I. G. A. S. (2022). The determinants of medical student learning behavior that are associated with the outcome of the Indonesia Medical Doctor National Competency Examination : A review. *Bali Medical Journal*, *11*(3), 2085–2089. https://doi.org/10.15562/bmj.v11i3.4426.

Darodjat, D., & Wahyudhiana, W. (2015). Model evaluasi program pendidikan. *Islamadina: Jurnal Pemikiran Islam*, *14*(1), 1–28. https://doi.org/10.30595/islamadina.v0i0.1665.

Daryono, R. W., Hariyanto, V. L., Usman, H., & Sutarto, S. (2020). Factor analysis: Competency framework for measuring student achievements of architectural engineering education in Indonesia. *REID (Research and Evaluation in Education)*, *6*(2), 98–108. https://doi.org/10.21831/reid.v6i2.32743.

Dellinges, M. A., & Curtis, D. A. (2017). Will a short training session improve multiple-choice item-writing quality by dental school faculty? A pilot study. *Journal of Dental Education*, *81*(8), 948–955. https://doi.org/10.21815/JDE.017.047.

Donnelly, C. (2014). The use of case based multiple choice questions for assessing large group teaching: implications on student’s learning. *Irish Journal of Academic Practice*, *3*(1), 1–15. https://doi.org/10.21427/D7CX32.

Elgadal, A. H., & Mariod, A. A. (2021). Item Analysis of Multiple-choice Questions (MCQs): Assessment Tool For Quality Assurance Measures. *Sudan Journal of Medical Sciences*, *16*(3), 334–346. https://doi.org/10.18502/sjms.v16i3.9695.

Gupta, P., Meena, P., Khan, A. M., Malhotra, R. K., & Singh, T. (2020). Effect of Faculty Training on Quality of Multiple-Choice Questions. *International Journal of Applied and Basic Medical Research*, *10*(3), 210–214. https://doi.org/10.4103/ijabmr.IJABMR_30_20.

Harden, R. M. (2018). Ten key features of the future medical school—not an impossible dream. *Medical Teacher*, *40*(10), 1010–1015. https://doi.org/10.1080/0142159X.2018.1498613.

Hijji, B. M. (2017). Flaws of multiple choice questions in teacher-constructed nursing examinations: A pilot descriptive study. *Journal of Nursing Education*, *56*(8), 490–496. https://doi.org/10.3928/01484834-20170712-08.

Kaur, M., Singla, S., & Mahajan, R. (2016). Item analysis of in use multiple choice questions in pharmacology. *International Journal of Applied and Basic Medical Research*, *6*(3), 170–173. https://doi.org/10.4103/2229-516X.186965.

Kibble, J. D. (2017). Best practices in summative assessment. *Advances in Physiology Education*, *41*(1), 110–119. https://doi.org/10.1152/advan.00116.2016.

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical Journal Armed Forces India*, *77*, S85–S89. https://doi.org/10.1016/j.mjafi.2020.11.007.

Kurtz, J. B., Lourie, M. A., Holman, E. E., Grob, K. L., & Monrad, S. U. (2019). Creating assessments as an active learning strategy: what are students' perceptions? A mixed methods study. *Medical Education Online*, *24*(1). https://doi.org/10.1080/10872981.2019.1630239.

León, S. P., Panadero, E., & García-Martínez, I. (2023). How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy. *Educational Psychology Review*, *35*(4), 106. https://doi.org/10.1007/s10648-023-09819-0.

Liew, C. P., Puteh, M., Mohammad, S., Omar, A. A., & Kiew, P. L. (2021). Review of engineering programme outcome assessment models. *European Journal of Engineering Education*, *46*(5), 834–848. https://doi.org/10.1080/03043797.2020.1852533.

Lockyer, J., Carraccio, C., Chan, M. K., Hart, D., Smee, S., & Touchie, C. (2017). Core principles of assessment in competency-based medical education. *Medical Teacher*, *39*(6), 609–616. https://doi.org/10.1080/0142159X.2017.1315082.

Mahda, A., Arfiyanti, M. P., Novitasari, A., & Romadhoni, R. (2023). Analisis Deskriptif Kualitas Soal Multiple Choice Questions (MCQ) Mini Kuis Tutorial di Fakultas Kedokteran Universitas Muhammadiyah Semarang. *Jurnal Ilmu Kedokteran Dan Kesehatan*, *10*(6), 2177–2184. https://doi.org/10.33024/jikk.v10i6.9932.

Mardiah, M., & Syarifudin, S. (2018). Model-Model Evaluasi Pendidikan. *MITRA ASH-SHIBYAN: Jurnal Pendidikan Dan Konseling*, *2*(1), 38–50. https://doi.org/10.46963/mash.v2i1.24.

Nojomi, M., & Mahmoudi, M. (2022). Assessment of multiple-choice questions by item analysis for medical students' examinations. *Research and Development in Medical Education*, *11*(1), 24. https://doi.org/10.34172/rdme.2022.024.

Przymuszała, P., Piotrowska, K., Lipski, D., Marciniak, R., & Cerbin-Koczorowska, M. (2020). Guidelines on Writing Multiple Choice Questions: A Well-Received and Effective Faculty Development Intervention. *SAGE Open*, *10*(3). https://doi.org/10.1177/2158244020947432.

Pugh, D., & Regehr, G. (2016). Taking the sting out of assessment: is there a role for progress testing? *Medical Education*, *50*(7), 721–729. https://doi.org/10.1111/medu.12985.

Rahma, N. A., Shamad, M. M., Idris, M. E., Elfaki, O. A., Elfakey, W. E., & Salih, K. M. (2017). Comparison in the quality of distractors in three and four options type of multiple choice questions. *Advances in Medical Education and Practice*, *8*, 287–291. https://doi.org/10.2147/AMEP.S128318.

Rodríguez, S. L. (2014). El aprendizaje basado en problemas para la educación médica: sus raíces epistemológicas y pedagógicas. *Revista Med*, *22*(2), 32–36. https://doi.org/10.18359/rmed.1168.

Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, *16*, 1–10. https://doi.org/10.1186/s12909-016-0773-3.

Shafira, N. N. A. (2015). Peran MCQ Sebagai Instrumen Evaluasi Dalam Pendidikan

Kedokteran. *Jambi Medical Journal: Jurnal Kedokteran Dan Kesehatan*, *3*(2), 132–139. https://doi.org/10.22437/jmj.v3i2.3089.

Tavakol, M., & Dennick, R. (2017). The foundations of measurement and assessment in medical education. *Medical Teacher*, *39*(10), 1010–1015. https://doi.org/10.1080/0142159X.2017.1359521.

Ten Cate, O., & Regehr, G. (2019). The Power of Subjectivity in the Assessment of Medical Trainees. *Academic Medicine*, *94*(3), 333–337. https://doi.org/10.1097/ACM.0000000000002495.

Utomo, P. S., Randita, A. B. T., Riskiyana, R., Kurniawan, F., Aras, I., Abrori, C., & Rahayu, G. R. (2022). Predicting medical graduates' clinical performance using national competency examination results in Indonesia. *BMC Medical Education*, *22*(1), 254. https://doi.org/10.1186/s12909-022-03321-x.

Xu, X., Kauer, S., & Tupy, S. (2016). Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Scholarship of Teaching and Learning in Psychology*, *2*(2), 147–158. https://doi.org/10.1037/stl0000062.

Zaidi, N. L. B., Grob, K. L., Monrad, S. M., Kurtz, J. B., Tai, A., Ahmed, A. Z., & Santen, S. A. (2018). Pushing Critical Thinking Skills With Multiple-Choice Questions: Does Bloom's Taxonomy Work? *Academic Medicine*, *93*(6), 856–859. https://doi.org/10.1097/ACM.0000000000002087.