Konakcı *BMC Medical Education* (2026) 26:327
https://doi.org/10.1186/s12909-026-08690-1

BMC Medical Education

RESEARCH Open Access

Equity in assessment: an examination of multiple-choice question tests

Serap Konakcı[1*]

Abstract

Background Equity and fairness are essential elements of acceptability, one of the Ottawa Criteria for good evaluation. The similarity of results from multiple-choice question (MCQ) tests developed under the same test blueprint for comparable groups can serve as empirical evidence of equity in assessment. This study aimed to examine whether MCQ tests prepared for repeated clerkship groups within one academic year demonstrated psychometric equivalence, focusing on item difficulty (DIF I) and discrimination (DI) indices as indicators of instrumental equity.

Methods Five MCQ tests administered at the end of the gynecology–urology internship during the 2022–2023 academic year were analyzed. All tests were developed in adherence to the same blueprint, ensuring identical content representation and cognitive level distribution. For each test, mean score, Kuder–Richardson 20 (KR-20) reliability coefficient, and standard error of measurement (SEM) were calculated. At the item level, difficulty index (DIF I), discrimination index (DI), and nonfunctional distractor (NFD) count were determined. Pairwise equivalence tests were conducted across five exams (ten comparisons) using 90% confidence intervals with predefined equivalence margins ($\Delta = \pm 0.10$). Equivalence was concluded when the entire CI of the mean difference fell within these limits.

Results Mean test scores ranged between 69 and 74; KR-20 values between 0.85 and 0.90; and SEM values between 3.44 and 3.64. While the mean scores differed significantly between groups ($F = 3.217$, $p = 0.013$, $\eta^2 = 0.039$), the differences were of low to moderate magnitude. All ten pairwise comparisons for both DIF I and DI met the predefined equivalence criterion, indicating psychometric similarity in item difficulty and discrimination across tests. However, none of the NFD comparisons met the equivalence criterion, suggesting minor fluctuations in distractor performance, likely due to technical variation rather than systematic bias.

Discussion and conclusion The fulfillment of equivalence criteria for DIF I and DI across all comparisons provides empirical evidence that the exams shared comparable psychometric properties. This consistency supports the notion of *instrumental equity*, reflecting fairness through the operational equivalence of assessment tools. Although full equivalence was not achieved for NFD rates, the overall psychometric consistency across tests indicates that student cohorts were evaluated with examinations of equivalent difficulty and discriminative capacity. Adhering to the test blueprint ensured balanced content representation, contributing to fairness and reliability in assessment. Routine monitoring of test-level and item-level metrics can serve as a practical approach to maintaining equity in medical education assessments.

*Correspondence:
Serap Konakcı
serapkonakci@yahoo.com

Full list of author information is available at the end of the article

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Keywords Instrumental equity, Multiple-choice questions, Item analysis, Psychometric equivalence

Introduction

Key criteria for a good assessment include validity or consistency, reproducibility, equivalence, applicability, educational impact, catalytic impact, and acceptability [1]. Assessment results affect a wide range of stakeholders [2]. Therefore, it is critically important that assessment methods and outcomes are acceptable to all stakeholders and accurately reflect reality. Unrealistic or biased results may lead to a waste of time and effort for students and inadequate health services for the community. This highlights the importance of *acceptability* and *fairness* in assessment [3, 4].

In the assessment process in medical education, it is crucial to ensure that students are not exposed to discrimination in any form and that *equity* and *justice* are maintained in evaluation. Lucey et al., described *equity in assessment* as a complex and multifaceted "wicked problem" [5]. Anderson et al., addressed the concept of fairness in assessment on a theoretical basis and proposed a three-dimensional framework comprising *fairness-oriented*, *assessment-for-inclusion*, and *justice-oriented* approaches [6]. Within this framework, *fairness-oriented assessment* emphasizes impartiality in decision-making, psychometric defensibility, and accountability. A subdimension of this approach, *instrumental equity*, focuses on how assessment data are used, interpreted, and compared across groups. In other words, the concept of fairness becomes operationalized through measurable equivalence. From this perspective, evaluating comparable student groups with examinations equivalent in structure, content, and psychometric characteristics can be regarded as an indicator of a fair and accountable assessment process.

Currently, multiple-choice question (MCQ) examinations are the most widely used tools for assessing knowledge in medical education worldwide [7–10]. MCQ analysis serves to evaluate student performance, compare groups taking the same exam, monitor exam and item quality, and assess program effectiveness. To ensure that MCQs are fit for purpose, their content must align with learning objectives [11]. Therefore, the use of *test blueprints* in exam preparation is recommended. Tests developed according to a blueprint not only enhance content validity but are also perceived by students as fairer [12].

Data obtained from standardized, blueprint-based examinations administered to comparable student groups are expected to be consistent. When testing conditions and item-writing standards are maintained equally, substantial differences between groups may indicate measurement inequity rather than true performance variation. Therefore, comparing test results provides a strong empirical basis for objectively examining equity in assessment.

In MCQ tests, descriptive statistics such as the mean, median, range, standard deviation (SD), standard error of the mean (SEM), and the Kuder–Richardson 20 (KR-20) coefficient are used to evaluate test-level characteristics [13–16]. At the item level, indices such as the Difficulty Index (DIF), Discrimination Index (DI), and Distractor Efficiency (DE) are used [14, 17–20]. DIF reflects the relative difficulty of an item, DI measures how well an item discriminates between high- and low-performing students, and the proportion of nonfunctional distractors (NFDs) indicates item design quality. Psychometric indicators such as KR-20, SEM, DIF, and DI can thus serve as quantifiable metrics for monitoring fairness in assessment.

In Türkiye, fairness and equity in assessment remain an underexplored and neglected topic in medical education. Despite the growing global emphasis on equity frameworks, empirical investigations examining fairness-oriented and instrumentally equitable assessment practices are still lacking in the Turkish context. Therefore, investigating *instrumental equity* based on assessment data represents an initial step toward translating theoretical discussions on equity into empirical inquiry. Examining whether multiple MCQ examinations, prepared under the same blueprint and administered to comparable student cohorts, are psychometrically equivalent provides a means of operationalizing fairness in assessment in an objective manner.

Methods

In this context, comparing examinations based on fundamental psychometric indicators is important for both testing equivalence and empirically demonstrating fairness. Accordingly, this study aims to provide measurable evidence supporting the concept of fairness in assessment within medical education.

The 5th semester program at Dokuz Eylül University Faculty of Medicine (DEUFM) consists of five blocks: skin-locomotor, emergency medicine, obstetrics and gynecology-urology, otorhinolaryngology-ophthalmology, and neurosciences. Students are assigned to small groups with similar characteristics at the start of the program, and these groups rotate through all blocks over the academic year. During the academic year, each block is repeated five times with different groups. In each block, students' performance is assessed through task evaluations conducted during the process, as well as through an MCQ test and an Observed Structured Clinical

Examination (OSCE) at the end. MCQ tests are used for knowledge assessment across all blocks.

This study compared data from five gynaecology-urology clerkship tests administered as part of the 5th semester programme during the 2022–2023 academic year. For each group, the tests were developed according to the test blueprint to ensure alignment with learning objectives, learning objective levels (according to Bloom' s taxonomy) and weighting in the block programme. To ensure content validity, the test blueprint was designed to ensure that each learning objective in the programme was tested by at least one MCQ [21]. When requesting MCQs from question writers, information about the relevant learning objective and objective level was provided, and question writers formulated their questions accordingly.

The tests comprised 120 five-choice, single-correct-answer MCQs. All questions in the five tests were newly created. Before the MCQs were included in the tests, they were reviewed by an evaluator for potential grammatical or logical clues, ensuring that correct options were not more detailed, all options were arranged chronologically or numerically, and that the stem did not contain unnecessary information.

Statistical analysis

The data for this research were obtained from the exam results and item analysis results retrieved from the education management system used in our faculty.

For each test, the mean score ± standard deviation (SD), minimum and maximum scores, median and range values, KR-20, and SEM were calculated. In addition, the distribution of the students according to the score groups was determined for each of the tests. Effect size was assessed using the Eta squared (η^2) value [22].

The DIF I and DF values for the test items were calculated as mean ± SD, along with their minimum and maximum values. A distractor with a preference rate of less than 5% was considered non-functional (NFD). The number of NFDs was recorded and calculated as mean ± SD for each MCQ.

To determine whether the examinations were psychometrically equivalent, pairwise equivalence analyses were performed among the five tests.

For each test pair, the mean values of the DIF I, DI and NFDs were compared.

Independent-samples t-tests were conducted, and 90% confidence intervals (CIs) were calculated for each comparison. Equivalence was established when the entire 90% CI of the mean difference fell within the predefined equivalence bounds ($\Delta = \pm 0.10$). This approach is mathematically equivalent to the Two One-Sided Tests (TOST) procedure [23–27].

All statistical analyses were conducted using SPSS 24.0 (IBM, Armonk, NY, United States of America).

The data used in the research were obtained retrospectively from the test scores, statistics and item analysis results of the examinations conducted by the Dean's Office of Dokuz Eylül University Faculty of Medicine. This research does not contain any participant data. The usage of the data has been granted permission by the Dean's Office of Dokuz Eylül University Faculty of Medicine. (14.09.2023/ Document number: E-45803587-755.02.01-726324).

The study was conducted after the approval of the Ethics Committee of Dokuz Eylül University for Non-interventional Studies was obtained (file no:8338-GOA, Decision No:2023/30 – 07).

Theoretical framework

Fairness in assessment has increasingly been conceptualized as a multidimensional construct shaped by psychometric, procedural, and sociocultural elements [4]. Within this landscape, fairness-oriented assessment frameworks emphasize impartiality in decision-making, transparency, defensibility, and the consistent use of evidence to guide judgments. Instrumental equity, a subdimension described in contemporary work on assessment equity [6], operationalizes fairness through measurable equivalence across assessment forms. In this model, fairness is not merely an ethical aspiration but a property demonstrable through observable psychometric indicators—such as item difficulty, discrimination, and reliability—which reflect whether different student groups are evaluated under comparable conditions.

By examining repeated MCQ examinations that were prepared under an identical test blueprint and administered to comparable student cohorts, this study aligns with theoretical perspectives that consider fairness an empirically testable dimension of assessment. This approach integrates the psychometric notion of equivalence with broader educational theories of justice, thereby positioning instrumental equity as a practical mechanism for evaluating fairness in medical education assessment.

Results

The number of students in the internship groups ranged from a minimum of 60 to a maximum of 67. The test with the largest range was the 2nd group's test, with a range of 52, while the test with the smallest range was the 5th group's test, with a range of 39. Across all tests, the mean scores varied between 69 and 74, KR-20 values ranged from 0.85 to 0.90, and SEM values were between 3.44 and 3.64 (Table 1).

When examining the distribution of students by score groups, it was observed that Test 2 had more students in the 55–59 score range than other tests, and fewer students in the 80–84 and 85–89 score ranges. The highest

Table 1 Characteristics of MCQ tests

Test no	No of student	Max. Score	Min. Score	Range	Mean	SD	Median	KR20	SEM
1	64	90	47	43	74	9.2	74	0.86	3.44
2	61	92	40	52	69	10.5	70	0.88	3.64
3	60	89	46	43	71	9.3	70.5	0.85	3.6
4	65	90	43	47	72	11.4	75	0.90	3.6
5	67	91	52	39	73	8.9	74	0.85	3.45

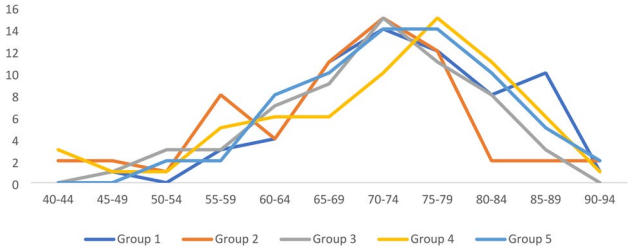


Fig. 1 Distribution of students by score groups

Table 2 DIF I and DI means of MCQs by the tests

	Test	Minimum	Maximum	Mean	SD
DIF I	1	.09	1.00	.73	.22
	2	.06	1.00	.67	.22
	3	.09	1.00	.71	.21
	4	.11	1.00	.70	.21
	5	.14	1.00	.73	.22
DI	1	−.12	.71	.23	.19
	2	−.19	.69	.25	.18
	3	−.12	.81	.23	.18
	4	−.33	.72	.27	.19
	5	−.11	.72	.22	.18

number of students scoring between 85 and 89 appeared in Test 1 (Fig. 1).

A comparison of test results based on student scores revealed statistically significant differences in mean test scores across the groups. However, the effect size analysis indicated that these differences were of low to moderate magnitude ($F = 3.217$, $p = 0.013$, and $\eta^2 = 0.039$).

An analysis of the item statistics showed that DIF I values ranged from very easy to very difficult, while DI values ranged from very low to good. The averages for DIF I and DI indicated that the exams were at acceptable levels. (Table 2).

The MCQs in the exams were categorized based on the number of non-functional distractors (NFDs). In the Group 1 exam, all distractors functioned properly in 1.7% of the MCQs, whereas in the Group 5 exam, this rate was 6.8%. MCQs with three non-functional distractors had the highest rate across all group tests (Table 3).

Pairwise equivalence analysis of the DIF I, DI and NFDs was conducted across five examinations, yielding ten pairwise comparisons for each parameter. All comparisons for both DIF I and DI satisfied the predefined equivalence criterion, as the entire 90% confidence intervals of the mean differences were contained within the ±0.10 bounds. In contrast, none of the ten comparisons for NFDs satisfied the equivalence criterion, as all 90% confidence intervals exceeded the ±0.10 bounds. (Table 4).

Discussion

In this study, we aimed to assess the equivalence of the exams by analyzing the results of the MCQ tests administered at the end of the obstetrics and gynecology/urology clerkship to five comparable student groups over the course of one academic year. The analysis revealed that the tests had similar characteristics in terms of measures of central tendency, score distribution, SEM, and KR-20 values. Although a statistically significant difference was found in the average scores between the groups, the differences were of low to moderate magnitude. All pairwise comparisons for both the Difficulty Index (DIF I) and the Discrimination Index (DI) met the predefined

Table 3 Distribution and means of MCQs in the tests according to the number of NFDs

Test		NFDs					Mean ± SD NFD
		0	1	2	3	4	
1	Count	2	16	22	48	29	2.27 ± 1.04
	%	1.7	13.7	18.8	41.0	24.8	
2	Count	6	24	28	41	21	2.39 ± 1.14
	%	5.0	20.0	23.3	34.2	17.5	
3	Count	6	21	27	42	24	2.48 ± 1.15
	%	5.0	17.5	22.5	35.0	20.0	
4	Count	7	18	34	39	21	2.41 ± 1.12
	%	5.9	15.1	28.6	32.8	17.6	
5	Count	8	21	26	31	31	2.48 ± 1.25
	%	6.8	17.9	22.2	26.5	26.5	

Table 4 Pairwise equivalence testing of DIF, DI and NFDs across five examinations

	Comparison	Mean difference (ΔM)	90% CI of ΔM	Equivalence bounds (±Δ)	Equivalent?
DIF	Exam 1–2	0.0620	[.01488, .01488]	±0.10	+
	Exam 1–3	0.0293	[-.01704, .07561]		+
	Exam 1–4	.03454	[-.01103, .08010]		+
	Exam 1–5	.00761	[-.03917, .05438]		+
	Exam 2–3	−.03267	[-.07919, .01385]		+
	Exam 2–4	−.02742	[-.07321, .01837]		+
	Exam 2–5	−.05435	[-.10133, .00737]		+
	Exam 3–4	.00525	[-.03987, .05028]		+
	Exam 3–5	−.02168	[-.06791, .02454]		+
	Exam 4–5	−.02693	[-.07239, .01854]		+
DI	Exam 1–2	−.02256	[-.06239, .01728]		+
	Exam 1–3	.00244	[-.03750, .04239]		+
	Exam 1–4	−.04323	[-.08416, −.00230]		+
	Exam 1–5	.00487	[-.03570, .04544]		+
	Exam 2–3	.02500	[-.01360, .06360]		+
	Exam 2–4	−.02068	[-.06026, .01891]		+
	Exam 2–5	.02743	[-.01178, .06663]		+
	Exam 3–4	−.04568	[-.08537, −.00598]		+
	Exam 3–5	.00243	[-.03689, .04174]		+
	Exam 4–5	.04810	[.00779, .08841]		+
NFDs	Exam 1–2	.142	[.109, .577]		-
	Exam 1–3	.260	[.026, .495]		-
	Exam 1–4	.323	[.091, .556]		-
	Exam 1–5	.256	[.008, .504]		-
	Exam 2–3	−.083	[-.327, .160]		-
	Exam 2–4	−.020	[-.262, .222]		-
	Exam 2–5	−.087	[-.343, .170]		-
	Exam 3–4	.063	[-.179; .305]		-
	Exam 3–5	−.004	[-.261, .253]		-
	Exam 4–5	−.067	[-.322, .188]		-

equivalence criterion. However, none of the ten comparisons for nonfunctional distractors (NFDs) met the equivalence criterion.

Designing the tests according to the blueprint ensured that all learning objectives were represented in each exam, with an equal distribution of the same objectives. This approach strengthened both the content validity and internal consistency of the tests. Moreover, the fact that the MCQs were developed by instructors actively involved in the clerkship program, and that detailed information about the learning objectives and their cognitive levels was shared with them, may have contributed to these consistent outcomes. The instructors were therefore aware of the topics and levels at which they needed to prepare MCQs.

Although our goal is always to obtain the most accurate results, measurement outcomes inevitably contain some degree of error. SEM values allow us to assess the extent of error in students' scores [13]. The SEM values across the tests were nearly identical (3.44, 3.64, 3.6, 3.6, and 3.45, respectively), indicating that measurement error was consistent for all groups, with no group appearing to have an advantage or disadvantage over others.

For end-of-course or end-of-year exams, a KR-20 value of 0.80 or above is generally considered acceptable [27, 28]. In our tests, KR-20 values ranged between 0.85 and 0.90, demonstrating high internal consistency across all tests, with very similar KR-20 values for each. Although the mean scores differed significantly between the groups ($F = 3.217$, $p = 0.013$, $\eta^2 = 0.039$), the differences were of low to moderate magnitude. This suggests that the variations likely stemmed from minor differences among the student groups rather than from the tests themselves. If the differences had been due to test design, other psychometric indicators would also have shown variation. Furthermore, since the assessment of internship success at our faculty includes multiple methods beyond a single MCQ test, this multimodal approach likely mitigates any limitations that could arise from the test design.

The fulfillment of equivalence criteria for both DIF I and DI across all pairwise comparisons provides strong empirical evidence that the examinations shared a comparable psychometric structure. This finding suggests

that each test consisted of items with similar levels of difficulty and discriminative power, administered under comparable testing conditions. From an educational perspective, this indicates that different student groups were evaluated through examinations of equivalent cognitive demand and measurement quality. Such consistency supports the notion of *instrumental equity*, which emphasizes fairness through the operational equivalence of assessment instruments.

Psychometric equivalence in item difficulty and discrimination is a crucial indicator of fairness-oriented assessment design. Downing [27] emphasized that the reproducibility and stability of psychometric properties across test forms are essential for maintaining fairness and defensibility in assessment. Similarly, Schuwirth and van der Vleuten [29] noted that equivalence between test versions ensures that performance differences reflect actual differences in competence rather than inconsistencies in measurement. Tavakol et al. [5] further highlighted that maintaining equivalent psychometric characteristics across groups demonstrates impartiality and accountability in assessment practices—core elements of fairness in medical education.

In the context of this study, achieving equivalence in DIF I and DI supports the inference that the blueprint-guided construction of examinations successfully ensured comparable item composition and psychometric quality. Consequently, it can be argued that the assessment process treated student cohorts equitably, upholding fairness not only conceptually but also empirically. These findings contribute to the growing discourse on fairness in medical education assessment, reinforcing the view that psychometric equivalence represents a measurable dimension of justice-oriented evaluation practices.

The NFD rate observed in our study was slightly higher than those reported in comparable studies [30–32]. However, full equivalence in the rate of NFDs across examinations was not achieved. This relatively higher NFD rate may be attributed to the fact that, unlike in some other studies, the MCQs used in this research included five options instead of four [33, 34]. Moreover, all MCQs were used for the first time and therefore had not undergone any post-hoc review process aimed at eliminating nonfunctioning distractors.

In addition, the acceptable mean DIF I values indicate that the influence of NFDs on item difficulty was not substantial [33]. The lack of equivalence in NFD counts likely reflects small, random variations in item-writing practices rather than systematic bias. Tarrant and Ware [35] emphasized that the efficiency of functional distractors is a major determinant of item quality and that a certain degree of variability between test administrations is to be expected. Similarly, Downing [36] noted that minor inconsistencies in distractor quality do not necessarily compromise test validity unless they occur systematically.

All MCQs in this study were newly developed and administered for the first time. Designing items that assess the same learning objectives and cognitive levels across five examinations within a single academic year poses a considerable challenge for item writers. Consequently, their primary focus may have been on formulating correct options rather than refining distractor quality. In this context, the observed inequalities in NFD rates are more likely to reflect minor technical differences and natural variation in item development processes rather than any compromise in the overall fairness of the assessment.

Our results demonstrate that the tests administered five times during one academic year in the internship program, while adhering to the test blueprint, produced comparable outcomes across all psychometric indicators evaluated. Based on measures of central tendency, SEM, KR-20, DIF I, and DI values, it can be concluded that equitable examinations were prepared for all groups. Adherence to the blueprint ensured that each test represented the same learning objectives with balanced cognitive coverage, thereby supporting both content validity and fairness in the evaluation process.

This study contributes to the literature by demonstrating how psychometric equivalence—particularly via DIF I and DI measures—can serve as empirical evidence of instrumental equity. While much of the fairness literature emphasizes conceptual definitions or structural inequities, this study provides a methodological example of how fairness can be operationalized and measured across repeated assessments. By linking theoretical perspectives on fairness with practical equivalence testing, the study supports the integration of psychometric analysis into justice-oriented assessment frameworks.

Limitations

This study has several limitations. First, the analysis relied entirely on psychometric data; perceptions of fairness among students and educators were not examined, limiting interpretive depth. Second, all MCQs were newly written and used for the first time without prior piloting, which may have influenced distractor functioning. Third, equivalence testing was conducted at the test level rather than at the individual item level, which may overlook finer-grained item performance patterns. Finally, as the study was conducted within a single institution and clerkship program, generalizability to other contexts may be limited.

Conclusion

Ensuring equity and fairness in evaluation is essential for maintaining both the acceptability and defensibility of assessments in medical education. This study

demonstrates that when multiple-choice examinations are developed under a clearly defined blueprint and key psychometric parameters—such as SEM, KR-20, DIF I, and DI—are systematically monitored, fairness can be empirically verified through *instrumental equity*.

The blueprint not only standardizes the test structure but also ensures comparable cognitive representation across exams, thereby promoting fairness through psychometric equivalence. Regular monitoring of test characteristics and item-level indicators strengthens internal quality assurance and provides a practical framework for ensuring equivalence across examinations.

Collectively, these practices help translate fairness in assessment from an abstract principle into a measurable, evidence-based dimension of educational evaluation.

Abbreviations

MCQ	Multiple-choice question
SD	Standard Deviation
SEM	Standard Error of the Mean
KR-20	Kuder-Richardson 20
DIF I	Difficulty Index
DI	Discrimination Index
DE	Distractor Efficiency
NFD	Nonfunctional Distractor
DEUFM	Dokuz Eylül University Faculty of Medicine
η^2	Eta squared

Organisational authorisation for data usage

The data used in the research were obtained retrospectively from the test scores, statistics and item analysis results of the examinations conducted by the Dean's Office of Dokuz Eylül University Faculty of Medicine. The usage of the data has been granted permission by the Dean's Office of Dokuz Eylül University Faculty of Medicine. (14.09.2023/ Document number: E-45803587-755.02.01-726324).

Authors' contributions

The author SK worked independently in conceptualizing and designing the study, collecting and analyzing data, drafting, and finalizing the manuscript.

Funding

No funding sources.

Data availability

The data used and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

The study was conducted after the approval of the Ethics Committee of Dokuz Eylül University for Non-interventional Studies was obtained (file no: 8338-GOA, Decision No: 2023/30 – 07) and was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki.

This research does not contain any participant data. It was conducted with statistical data from the MCQ tests. Therefore, it is NOT APPLICABLE /NOT POSSIBLE to obtain 'informed consent from participants'.

Consent for publication

Not applicable.

This study was conducted by analysing retrospectively obtained data onMCQ test statistics. does not contain any personal or clinical details ofindividuals in any way.

Competing interests

The authors declare no competing interests.

Author details

[1]Faculty of Medicine Medical Education Department, Dokuz Eylül University, İzmĩr, Türkiye

Received: 2 October 2024 / Accepted: 22 January 2026

Published online: 29 January 2026

References

1. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206 – 14. https://doi.org/10.3109/0142159X.2011.551559 PMID: 21345060.
2. Amin Z, Seng CY, Eng KH. Practical guide to medical student assessment. Singapore: World Scientific Publishing Co. Ltd. Pte; 2006.
3. Singh T. Principles of assessment in medical education. New Delhi: Jaypee Brothers Medical; 2021.
4. Lucey CR, Hauer KE, Boatright D, Fernandez A. Medical education's wicked problem: achieving equity in assessment for medical learners. Acad Med. 2020;95(12 Suppl):S98–108. https://doi.org/10.1097/ACM.0000000000003717.
5. Tavakol M, Stewart C, Sharpe CC. Ensuring fairness in assessment in health professions education: rapid analysis tools to detect differential item functioning across groups. Int J Med Educ. 2024;15:80–3. https://doi.org/10.5116/ijme.6694.de69.
6. Anderson HLK, Abdulla L, Balmer D, Govaerts M, Busari JO, West DC. Clarifying and expanding equity in assessment by considering three orientations: Fairness, inclusion, and justice. Med Educ. 2024. https://doi.org/10.1111/medu.15534. [Epub ahead of print].
7. Coderre SP, Harasym P, Mandin H, Fick G. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. BMC Med Educ. 2004;4:23. https://doi.org/10.1186/1472-6920-4-23.
8. Gierl MJ, Lai H, Turner SR. Using automatic item generation to create multiple-choice test items. Med Educ. 2012;46(8):757–65. https://doi.org/10.1111/j.1365-2923.2012.04289.x.
9. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. BMC Med Educ. 2007;7:49. https://doi.org/10.1186/1472-6920-7-49.
10. 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(2):198–206. https://doi.org/10.1111/j.1365-2923.2007.02957.x.
11. Gottlieb M, Bailitz J, Fix M, Shappell E, Wagner MJ. Educator's blueprint: A how-to guide for developing high-quality multiple-choice questions. AEM Educ Train. 2023;7(1):e10836. https://doi.org/10.1002/aet2.10836. PMID: 36711253; PMCID: PMC9873868.
12. Patil SY, Gosavi M, Bannur HB, Ratnakar A. Blueprinting in assessment: A tool to increase the validity of undergraduate written examinations in pathology. Int J Appl Basic Med Res. 2015;5(1):S76–9. https://doi.org/10.4103/2229-516X.162286. PMID: 26380218; PMCID: PMC4552073.
13. Tavakol M, O'Brien D. Psychometrics for physicians: everything a clinician needs to know about assessments in medical education. Int J Med Educ 2022 Apr: 22 (13):100–6. https://doi.org/10.5116/ijme.625f.bfb1. PMID: 35462355; PMCID: PMC9902176.
14. Tavakol M, Dennick R. Post-examination analysis of objective tests. Medical Teacher. 2011; 33(6): 447–458. Available from: Available from: https://doi.org/10.3109/0142159X.2011.564682
15. Tavakol M, Dennick R. Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE guide 66. Med Teach. 2012;34(3):e161–75.
16. Kuder GF, Richardson MW. The theory of the Estimation of test reliability. Psychometrika. 1937;2:151–60. https://doi.org/10.1007/BF02288391.
17. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. 3rd ed. Philadelphia: National Board of Medical Examiners; 2000. Available from: https://health.uconn.edu/faculty-development/wp-content/uploads/sites/69/2017/06/constructing_written_test_questions.pdf
18. Medical Council of Canada (MCC). Guidelines for the Development of Multiple-Choice Questions. Ottawa, ON: MCC 2010. Available from: https://mcc.ca/wp-content/uploads/Multiple-choice-question-guidelines.pdf
19. Date AP, Borkar AS, Badwaik RT, Siddiqui RA, Shende TR, Dashputra AV. Item analysis as tool to validate multiple choice question bank in pharmacology.

International Journal of Basic & Clinical Pharmacology. 2019;8(9):1999–2003. Available from: https://doi.org/10.18203/2319-2003.ijbcp20194106
20. Rao C, Kishan Prasad HL, Sajitha K, Permi H, Shetty J. Item analysis of multiple choice questions: Assessing an assessment tool in medical students. Int J Educ Psychol Res [serial online]2016;2:201-4. Available from: https://www.ijeprjournal.org/text.asp?2016/2/4/201/189670
21. Bridge PD, Musial J, Frank R, Roe T, Sawilowsky S. Measurement practices: methods for developing content-valid student examinations. Med Teach. 2003;25(4):414–21.
22. Cohen J. Statistical Power Analysis for the Behavioral Sciences. New, York NY. Routledge Academic; 1988. Available from: https://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf
23. Schuirmann DJ. A comparison of the two One-Sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1987;15(6):657–80. https://doi.org/10.1007/BF01068419.
24. Lakens D. Equivalence tests: a practical primer for t-tests, correlations, and meta-analyses. Soc Psychol Pers Sci. 2017;8(4):355–62. https://doi.org/10.1177/1948550617697177.
25. Jodoin MG, Gierl MJ. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Appl Meas Educ. 2001;14(4):329–49. https://doi.org/10.1207/S15324818AME1404_2.
26. Raju NS, van der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Appl Psychol Meas. 1995;19(4):353–68. https://doi.org/10.1177/014662169501900405.
27. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38(9):1006–12.
28. Belay LM, Sendekie TY, Eyowas FA. Quality of multiple-choice questions in medical internship qualification examination determined by item response theory at Debre Tabor University, Ethiopia. BMC Medical Education. 2022; 22(1):635. Available from: https://doi.org/10.1186/s12909-022-03687-y
29. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478–485. Available from: https://doi.org/10.3109/0142159X.2011.565828
30. Tarrant M, Ware J. A framework for improving the quality of multiple-choice assessments. Nurse Educ. 2012 May-Jun;37(3):98–104. https://doi.org/10.1097/NNE.0b013e31825041d0. PMID: 22513766.
31. 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(2):133 – 43. https://doi.org/10.1007/s10459-004-4019-5. PMID: 16078098.
32. Bhattacherjee S, Mukherjee A, Bhandari K, Rout AJ. Evaluation of Multiple-Choice questions by item Analysis, from an online internal assessment of 6th semester medical students in a rural medical College, West Bengal. Indian J Community Med. 2022;47(1):92–5. https://doi.org/10.4103/ijcm.ijcm_1156_21.
33. Hingorjo MR, Jaleel F. Analysis of one-best mcqs: the difficulty index, discrimination index and distractor efficiency. J Pak Med Assoc. 2012;62(2):142–7.
34. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Medical Education, 2009; 9: 40. Available from: https://doi.org/10.1186/1472-6920-9-40
35. Al-lawama M, Kumwenda B. Decreasing the options' number in multiple choice questions in the assessment of senior medical students and its effect on exam psychometrics and distractors' function. BMC Med Educ. 2023;23(1):212.
36. Rodriguez MC. Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. Educational Measurement: Issues and Practice. 2005; 24(2):3–13. Available from: https://doi.org/10.1111/j.1745-3992.2005.00006.x

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.