

Impact d'une formation à la rédaction des QCM sur la qualité psychométrique d'une épreuve de chirurgie: Etude avant-après

Impact of MCQ-Writing Training on the Psychometric Quality of a Surgery Examination: A Before-and-After Study

Rayen Lahouar^{1,2}, Baha Khelifa^{1,2}, Jihène Krichène^{1,3}, Mohamed Kahloul^{1,4}, Badreddine Ben Khalifa^{1,2}, Wael Gazzah^{1,2}, Sahbi Naouar^{1,2}, Braiek Salem^{1,2}

1. Faculté de médecine Ibn El Jazzar Sousse, Rue Mohamed Karoui, 4002 Sousse, Tunisie
2. Hôpital Universitaire Ibn El Jazzar de Kairouan, Service d'urologie, Rue Ibn El Jazzar, 3100 Kairouan, Tunisie
3. Hôpital Universitaire Ibn El Jazzar de Kairouan, Service de chirurgie générale, Rue Ibn El Jazzar, 3100 Kairouan, Tunisie
4. Hôpital universitaire Sahloul, Service d'anesthésiologie et de réanimation, 4011, Sousse, Tunisie.

RÉSUMÉ

Introduction : L'évaluation sommative joue un rôle central dans la formation médicale et nécessite des outils valides et fiables. La qualité psychométrique des questions à choix multiples (QCM) dépend largement de leur conception.

Objectif : Évaluer l'impact d'une formation pédagogique à la rédaction de QCM sur les propriétés psychométriques d'un examen de chirurgie à la Faculté de médecine de Sousse.

Méthodes : Il s'agit d'une étude pré-expérimentale avant-après comparant deux sessions successives du même examen, menées en 2024 et 2025. Au total, 174 étudiants (96 en période 1 et 78 en période 2) et 180 items ont été analysés. Les paramètres évalués comprenaient l'indice de difficulté, l'indice de discrimination, le coefficient d'efficacité, les catégories de réponse et les indicateurs globaux du test, notamment l'alpha de Cronbach et l'erreur de mesure.

Résultats : Les performances globales des étudiants, en termes de score moyen et de taux de réussite, étaient comparables entre les deux périodes ($p = 0,774$). Cependant, l'indice de discrimination médian s'est significativement amélioré, passant de 28,27 % à 40 % ($p = 0,003$). La proportion d'items présentant une excellente discrimination a augmenté de 23,3 % à 50 % ($p < 0,001$). Le coefficient alpha de Cronbach a progressé de 0,87 à 0,93, tandis que l'indice de difficulté est resté stable.

Conclusion : La formation pédagogique a amélioré la qualité psychométrique de l'examen sans modifier les performances globales des étudiants, soulignant ainsi la valeur d'une approche institutionnelle fondée sur la formation continue du corps professoral et l'analyse psychométrique régulière des examens.

Mots clés : Évaluation de certification, questions à choix multiples, l'analyse des articles, indice de discrimination, développement du corps professoral, formation médicale

ABSTRACT

Introduction: Summative assessment plays a central role in medical education and requires valid and reliable tools. The psychometric quality of multiple-choice questions (MCQs) largely depends on their design.

Objective: To assess the impact of pedagogical training in MCQ writing on the psychometric properties of a surgery examination at the Faculty of Medicine of Sousse.

Methods: This was a pre-experimental before-and-after study comparing two successive sessions of the same examination, conducted in 2024 and 2025. A total of 174 students, including 96 in period 1 and 78 in period 2, as well as 180 items, were analyzed. The assessed parameters included the difficulty index, discrimination index, effective coefficient, response categories, and global test indicators, particularly Cronbach's alpha and measurement error.

Results: Overall student performance, in terms of mean score and success rate, was comparable between the two periods ($p = 0.774$). However, the median discrimination index significantly improved, increasing from 28.27 to 40% ($p = 0.003$). The proportion of items with excellent discrimination increased from 23.3 to 50% ($p < 0.001$). Cronbach's alpha increased from 0.87 to 0.93, while the difficulty index remained stable.

Conclusion: Pedagogical training improved the psychometric quality of the examination without changing overall student performance, highlighting the value of an institutional approach based on continuous faculty training and regular psychometric analysis of examinations.

Keywords: Certifying assessment; multiple-choice questions; item analysis; discrimination index; faculty development; medical education.

Correspondance

Baha Khelifa

Hôpital Universitaire Ibn El Jazzar de Kairouan, Service d'urologie, Rue Ibn El Jazzar, 3100 Kairouan, Tunisie

Email: khelifabaha@gmail.com

INTRODUCTION

L'évaluation certificative occupe une place centrale dans la formation médicale, car elle influence les apprentissages et conditionne la crédibilité des décisions académiques. Dans ce contexte, l'enjeu n'est plus seulement de tester les connaissances, mais de garantir des épreuves valides, fiables, équitables et alignées sur les compétences attendues (1). Les questions à choix multiples (QCM) restent très utilisées en raison de leur objectivité, de leur faisabilité et de leur capacité à couvrir un large contenu pédagogique (2).

Cependant, la qualité psychométrique d'une épreuve dépend fortement de la conception des items. Les ambiguïtés rédactionnelles, l'absence de blue-print, les distracteurs peu fonctionnels et les défauts d'alignement peuvent altérer la fidélité, la validité et le pouvoir discriminant de l'examen (3,4). L'analyse docimologique, fondée sur l'indice de facilité, l'indice de discrimination, la cohérence interne et l'analyse des réponses, constitue ainsi un outil d'assurance qualité des examens (2-4). Pour cette raison, la formation des enseignants, notamment sur les principes d'évaluation et la rédaction des QCM, est souvent proposée comme levier d'amélioration (5).

À la Faculté de Médecine de Sousse, une analyse docimologique réalisée en 2024 sur l'épreuve écrite du mi-bloc de chirurgie en 1ère année du deuxième cycle d'enseignement médical (DCEM1) a mis en évidence des indicateurs décevants par rapport aux résultats attendus. Dans ce contexte, une formation pédagogique des enseignants, centrée sur les règles de confection et de validation des QCM, a été organisée en janvier 2025. L'objectif de ce travail était de comparer les propriétés psychométriques de cette épreuve avant et après cette formation. L'hypothèse était que cette intervention améliorerait la qualité des items et la fidélité globale de l'épreuve, sans nécessairement modifier les performances globales des étudiants.

METHODES

Il s'agissait d'une étude pré-expérimentale type avant-après portant sur deux versions successives de l'examen écrit du mi-bloc de chirurgie de DCEM1 à la Faculté de Médecine de Sousse. La période 1 (P1) correspondait à la session de 2024, avant la formation pédagogique des enseignants des disciplines du bloc de chirurgie, et la période 2 (P2) à la session de 2025, après cette formation. Il s'agissait d'un séminaire pédagogique portant sur les principes de l'évaluation, la rédaction des QCM, la validation des questions et l'interprétation des indicateurs docimologiques.

Population de l'étude

La population étudiée comprenait tous les étudiants ayant passé l'examen du mi-bloc de chirurgie durant les deux périodes : 96 étudiants en P1 et 78 en P2. Chaque épreuve comportait 90 items de type QCM couvrant les disciplines chirurgicales enseignées, soit 180 items

analysés au total.

Les données ont été extraites des fichiers de notes et des rapports statistiques de la plateforme d'évaluation. Les résultats ont été exploités sous forme anonymisée. L'unité d'analyse principale était l'item.

Variables étudiées

L'analyse globale a porté sur l'effectif, la note moyenne sur 20, les extrêmes, la médiane, l'intervalle interquartile et le taux de réussite, défini par une note supérieure ou égale à 10/20.

L'analyse par item a inclus l'indice de facilité, l'indice de discrimination, le coefficient effectif et les fréquences de réponses correctes, partiellement correctes et incorrectes. Les réponses partiellement correctes correspondaient aux réponses recevant une note intermédiaire entre zéro et la note maximale, selon une logique de crédit partiel (6,7). On s'est intéressé particulièrement à étudier les questions dont la proportion de réponse incorrecte > 30% et à celles dont la proportion de réponse correcte > 80%.

L'indice de facilité était interprété comme faible en dessous de 30 %, acceptable entre 30 % et 70 %, et élevé au-dessus de 70 %. Pour l'indice de discrimination, les catégories retenues étaient : négative (<0), faible (0-19,99 %), acceptable (20-29,99 %), bonne (30-39,99 %) et excellente (≥40 %).

L'analyse par épreuve comprenait la dissymétrie (Skewness), l'aplatissement (Kurtosis), l'alpha de Cronbach, l'erreur de rapport et l'erreur standard. La dissymétrie et l'aplatissement ont été utilisés pour apprécier la structure de la distribution des notes (8). L'alpha de Cronbach a été interprété comme un indicateur de cohérence interne (9). L'erreur standard renseignait sur la précision de la mesure individuelle (10). L'analyse des distracteurs n'a pas été réalisée de manière fiable, faute de données détaillées option par option.

Analyse statistique

Les variables qualitatives ont été décrites par effectifs et pourcentages. Les variables quantitatives ont été présentées en moyennes ± écart-type ou en médianes avec intervalles interquartiles selon la distribution. Les comparaisons ont utilisé le test du chi-deux de Pearson ou le test exact de Fisher pour les variables qualitatives, le test t de Student ou l'ANOVA pour les moyennes, et le test de Mann-Whitney pour les médianes. Le seuil de significativité était fixé à 5 %. Les paramètres disponibles une seule fois par épreuve, tels que l'alpha de Cronbach ou l'erreur de rapport, ont été comparés de manière descriptive. Les analyses ont été réalisées avec le logiciel IBM SPSS dans sa version 25.0.

Considérations éthiques

Cette étude ne modifiait en rien les décisions académiques prises vis-à-vis des étudiants. Elle reposait sur des données déjà produites, préalablement saisies sur fichiers Excel et analysées secondairement sous forme anonymisée.

RÉSULTATS

Au total, 174 copies ont été analysées, correspondant à 96 étudiants en P1 et 78 étudiants en P2. L'analyse docimologique a porté sur 180 items, soit 90 items par période. Toutes les notes ont été exprimées sur 20.

Performances globales

Les performances globales des étudiants étaient comparables entre les deux périodes. La note moyenne était de $8,99 \pm 1,92$ en P1 contre $9,22 \pm 2,58$ en P2, sans différence significative. Le taux de réussite a légèrement augmenté de 36,5 % à 39,7 %, sans différence statistiquement significative.

Tableau 1. Performances globales des étudiants selon la période

Paramètre	P1	P2	p
Effectif	96	78	-
Note moyenne /20 $\pm$ ET	$8,99 \pm 1,92$	$9,22 \pm 2,58$	0,525
Extrêmes	3,27-13	4,34-13,76	-
Taux de réussite n (%)	35 (36,5)	31 (39,7)	0,774

ET : écart-type

Paramètres psychométriques par item

L'indice de facilité moyen était proche entre P1 et P2. En revanche, l'indice global de discrimination était significativement plus élevé après la formation. Le coefficient effectif et les fréquences de réponses correctes, partiellement correctes et incorrectes ne différaient pas significativement entre les deux périodes (Tableau 2).

Tableau 2. Comparaison des paramètres psychométriques par item

Paramètres	P1	P2	p
Indice de facilité, moyenne $\pm$ ET (min-max)	$44,93 \pm 19,20$ (1,74-90,97)	$46,11 \pm 16,38$ (5,13-94,87)	0,659
Indice global de discrimination, médiane [Q1-Q3]	28,27 [13,69-39,40]	40 [13,57-53,62]	0,003
Coefficient effectif, moyenne $\pm$ ET (min-max)	$1,20 \pm 0,37$ (0,22-1,89)	$1,23 \pm 0,37$ (0,27-1,97)	0,545
Items à covariance négative, n (%)	7 (7,8)	9 (10)	-
Taux de Réponses correctes, médiane en % [Q1-Q3]	13,50 [5-33]	16 [8-36,50]	0,407
Taux de Réponses partiellement correctes, médiane en % [Q1-Q3]	40 [16,50-71]	49 [6,50-70,50]	0,994
Taux de Réponses incorrectes, médiane en % [Q1-Q3]	28,50 [14-52,75]	28 [14-46,75]	0,723

ET : écart-type ; Q1-Q3 : intervalle interquartile 25-75.

Analyse selon les seuils docimologiques

La répartition des items selon les seuils de facilité ne différait pas significativement entre les deux périodes. En revanche, la distribution globale de l'indice de discrimination montrait une différence significative, marquée par une augmentation nette des items à discrimination excellente en P2 (23,3 % versus 50% dans P1 et P2 respectivement ; $p < 0,001$). Parallèlement, la proportion des items à bonne discrimination a significativement diminué, passant de 22,2 % en P1 à 10 % en P2 ($p = 0,043$).

Les proportions d'items avec au moins 80 % de réponses correctes et d'items avec au moins 30 % de réponses incorrectes étaient comparables (Tableau 3).

Tableau 3. Répartition des items selon les seuils docimologiques

Indicateur	P1	P2	p
Indice de facilité global, moyenne $\pm$ ET	$44,93 \pm 19,20$	$46,11 \pm 16,38$	0,725
Items difficiles, n (%)	23 (25,6)	13 (14,4)	0,062
Items de difficulté acceptable, 60 n (%)	66 (66,7)	71 (78,9)	0,065
Items faciles, n (%)	7 (7,8)	6 (6,7)	0,773
Indice de discrimination global, médiane [Q1-Q3]	28,27 [13,69-39,40]	40 [13,57-53,62]	0,003
Items à discrimination excellente, n (%)	21 (23,3)	45 (50)	<0,001
Items à discrimination bonne, n (%)	20 (22,2)	9 (10)	0,043
Items à discrimination acceptable, n (%)	18 (20)	9 (10)	0,095
Items à faible discrimination, n (%)	21 (23,3)	17 (18,9)	0,584
Items à discrimination négative, n (%)	10 (11,1)	10 (11,1)	1
Taux de réponses correctes $\geq$ 80 %, n (%)	4 (4,4)	2 (2,2)	0,682
Taux de réponses incorrectes $\geq$ 30 %, n (%)	44 (48,9)	43 (47,8)	0,881

ET : écart-type ; Q1-Q3 : intervalle interquartile.

Analyse par épreuve

À l'échelle de l'épreuve, la cohérence interne était élevée dans les deux périodes, avec une amélioration descriptive de l'alpha de Cronbach en P2. L'erreur de rapport a diminué, alors que l'erreur standard est restée stable. Ces paramètres n'ont pas fait l'objet d'un test inférentiel, car ils n'étaient disponibles qu'une seule fois par période.

Tableau 4. Paramètres docimologiques globaux par épreuve

Paramètre	P1	P2	p
Dissymétrie	-0,33	-0,10	NA
Aplatissement	-0,32	-0,97	NA
Alpha de Cronbach	0,87	0,93	NA
Erreur de rapport	36,1 %	26,5 %	NA
Erreur standard	0,69	0,68	NA

NA : non applicable.

DISCUSSION

À notre connaissance, les données maghrébines évaluant l'impact d'une formation pédagogique des enseignants aux règles de confection des examens sur les indicateurs psychométriques d'une épreuve médicale réelle restent limitées. Dans ce contexte, notre étude constitue l'un des premiers travaux à explorer cet effet dans un cadre certificatif authentique. Elle a montré que la formation pédagogique des enseignants n'a pas entraîné de modification significative des performances globales des étudiants. La note moyenne ainsi que le taux de réussite sont restés comparables entre les deux périodes. En revanche, plusieurs indicateurs de qualité psychométrique se sont améliorés, notamment l'indice de discrimination, la proportion d'items à discrimination excellente et la cohérence interne de l'épreuve. L'effet observé semble donc concerner davantage la qualité de mesure de l'examen que son niveau apparent de difficulté.

Cette distinction mérite d'être soulignée. En effet, une formation à la rédaction des QCM n'a pas pour objectif de rendre l'épreuve plus facile, mais plutôt de la rendre plus équitable, plus fiable et plus apte à différencier les étudiants selon leur niveau réel de maîtrise. Dans notre série, la stabilité de l'indice de facilité suggère que le niveau global d'exigence de l'épreuve a été maintenu. Ce résultat rejoint les données de la littérature, qui soulignent que la qualité d'un examen repose principalement sur une majorité d'items de difficulté intermédiaire, associée à une réduction des items extrêmes, trop faciles ou trop difficiles (3,11,12).

Le résultat le plus marquant de notre étude concerne l'amélioration significative de l'indice global de discrimination, dont la médiane est passée de 28,27 % à 40 % après la formation. Parallèlement, la proportion d'items à discrimination excellente a doublé, passant de 23,3 % à 50 %. À l'inverse, celle des items à bonne discrimination a diminué de manière significative. Cette baisse pourrait s'expliquer par le passage d'une partie des items initialement classés comme ayant une bonne discrimination vers la catégorie des items à discrimination excellente. Dans l'ensemble, cette évolution traduit une amélioration réelle de la qualité discriminante de l'épreuve.

Ces résultats sont en accord avec les données de la littérature soulignant l'intérêt des programmes de développement pédagogique dans l'amélioration de la qualité des QCM, en particulier lorsqu'ils sont réguliers, contextualisés et accompagnés d'un feedback psychométrique structuré (13–15). À l'inverse, les formations ponctuelles, courtes ou isolées semblent avoir un impact plus limité, parfois inconstant, comme l'ont rapporté Gupta et al. ainsi que Dellings et Curtis (16,17). Malgré cette amélioration, 11,1 % des items conservaient une discrimination négative dans les deux périodes. Ces items doivent être considérés comme prioritaires pour la révision, car ils peuvent refléter une ambiguïté, une erreur de clé, un distracteur inadéquat ou un défaut d'alignement pédagogique. Les défauts de rédaction des QCM sont connus pour altérer la qualité psychométrique des examens et justifient une relecture collégiale avant

l'épreuve ainsi qu'une analyse systématique après l'examen (4,18,19).

L'alpha de Cronbach est passé de 0,87 à 0,93, traduisant une amélioration descriptive de la cohérence interne de l'épreuve. Ces valeurs apparaissent satisfaisantes et sont comparables, voire supérieures, à celles rapportées dans plusieurs études docimologiques menées en contexte médical (4,20). La diminution de l'erreur de mesure va également dans le sens d'une meilleure fiabilité de l'examen. Toutefois, ces indicateurs n'ayant été disponibles qu'une seule fois pour chaque période, leur interprétation doit rester prudente.

Par ailleurs, l'absence d'analyse détaillée des distracteurs constitue une limite importante. Plusieurs travaux ont montré que les distracteurs non fonctionnels influencent directement l'indice de facilité et le pouvoir discriminant des items (21). Rezigalla et al. ont ainsi montré que l'efficacité des distracteurs avait un impact significatif sur l'indice de difficulté et sur le pouvoir discriminant des QCM (22). Rashwan et al., à leur tour, ont également souligné qu'un grand nombre de distracteurs non fonctionnels pourrait justifier la révision d'un item plutôt que son exclusion pure et simple (4). Dans notre contexte, l'accès aux réponses option par option permettrait d'expliquer plus finement les variations observées et de construire une banque locale d'items validés.

Nos résultats soutiennent l'idée qu'une formation pédagogique des enseignants, lorsqu'elle est associée à une analyse docimologique régulière, peut améliorer la qualité psychométrique des épreuves sans nécessairement modifier les performances globales des étudiants. Cette approche s'inscrit dans une démarche d'assurance qualité des évaluations, où l'objectif n'est pas de rendre l'examen plus facile, mais de le rendre plus valide, plus fiable, plus équitable et plus discriminant (23). Les données récentes de la littérature confirment cet intérêt. La revue systématique de Hye et al. a montré que les interventions visant à améliorer les QCM, notamment la formation des enseignants et la relecture par les pairs, améliorent la qualité des items, en particulier lorsque ces interventions sont répétées, structurées et intégrées dans une démarche institutionnelle de contrôle qualité. Les ateliers isolés semblent avoir un effet plus limité, alors que les programmes longitudinaux et le feedback régulier à partir des indices psychométriques produisent des améliorations plus durables (5).

En pratique, l'intérêt principal des formations pédagogiques est d'aider les enseignants à mieux concevoir leurs QCM : formuler des énoncés plus clairs, mieux aligner les questions avec les objectifs pédagogiques, réduire les ambiguïtés et améliorer la qualité des distracteurs. Toutefois, la formation ne peut pas, à elle seule, garantir la qualité durable des examens. Elle doit être intégrée dans une démarche continue associant un « blue-printing » explicite, relecture des items par les pairs, une analyse psychométrique après chaque session, un feedback aux enseignants et une révision des questions problématiques. Cette démarche rejoint les données tunisiennes récentes de Keskes et al. qui, dans une analyse docimologique d'une épreuve d'ECOS, soulignent l'intérêt de l'exploitation régulière

des indicateurs docimologiques, de l'optimisation des outils d'évaluation et du retour systématique des résultats aux évaluateurs. Cette boucle continue permettrait d'améliorer progressivement la validité, la fidélité, la discrimination et l'équité des examens (24–27).

Notre étude présente certaines forces. Tout d'abord, elle porte sur une épreuve réelle à enjeu certificatif, inclut l'ensemble des étudiants de chaque période et analyse les items un par un. Elle apporte aussi des données locales sur l'impact mesurable d'une formation pédagogique dans un contexte maghrébin où ce type d'étude reste rare. Néanmoins, ses limites sont liées au caractère rétrospectif, monocentrique, à l'absence de groupe contrôle, aux différences possibles entre cohortes et contenus d'épreuve, à l'absence de « blue-print » détaillé et à l'impossibilité d'analyser les distracteurs et les défauts rédactionnels.

En conclusion, une formation pédagogique des enseignants sur les principes de confection des QCM a été associée à une amélioration de la qualité psychométrique d'une épreuve de chirurgie en DCEM1, principalement en termes de discrimination des items et de cohérence interne, sans modification significative des performances globales des étudiants. Cette amélioration reste partielle et justifie la mise en place d'une stratégie institutionnelle durable associant une formation continue, relecture collégiale, analyse docimologique régulière et constitution progressive d'une banque d'items validés.

RÉFÉRENCES

- Kassam A, de Vries I, Zabar S, Durning SJ, Holmboe E, Hodges B, et al. The Next Era of Assessment Within Medical Education: Exploring Intersections of Context and Implementation. *Perspect Med Educ.* 2024;13(1):496-506. doi:10.5334/pme.1128 PubMed PMID: 39399409; PubMed Central PMCID: PMC11469546.
- Zubairi NA, AlAhmadi TS, Ibrahim MH, Hegazi MA, Gadi FU. Effective use of Item Analysis to improve the Reliability and Validity of Undergraduate Medical Examinations: Evaluating the same exam over many years: a different approach. *Pak J Med Sci.* mars 2025;41(3):810-5. doi:10.12669/pjms.41.3.10693 PubMed PMID: 40103890; PubMed Central PMCID: PMC11911747.
- Eleragi AMS, Miskeen E, Hussein K, Rezigalla AA, Adam MIE, Al-Faifi JA, et al. Evaluating the multiple-choice questions quality at the College of Medicine, University of Bisha, Saudi Arabia: a three-year experience. *BMC Med Educ.* 13 févr 2025;25(1):233. doi:10.1186/s12909-025-06700-2 PubMed PMID: 39948528; PubMed Central PMCID: PMC11827222.
- Rashwan NI, Aref SR, Nayel OA, Rizk MH. Postexamination item analysis of undergraduate pediatric multiple-choice questions exam: implications for developing a validated question Bank. *BMC Med Educ.* 21 févr 2024;24(1):168. doi:10.1186/s12909-024-05153-3 PubMed PMID: 38383427; PubMed Central PMCID: PMC10882907.
- Hye T, Hossain MF, Roni MA. Interventions to Improve the Quality of Multiple-Choice Questions: A Systematic Review and Meta-Analysis. *Am J Pharm Educ.* nov 2025;89(11):101872. doi:10.1016/j.ajpe.2025.101872 PubMed PMID: 40962033.
- Betts J, Muntean W, Kim D, Kao S chuan. Evaluating Different Scoring Methods for Multiple Response Items Providing Partial Credit. *Educ Psychol Meas.* févr 2022;82(1):151-76. doi:10.1177/0013164421994636 PubMed PMID: 34992310; PubMed Central PMCID: PMC8725057.
- Masters GN. A rasch model for partial credit scoring. *Psychometrika.* 1 juin 1982;47(2):149-74. doi:10.1007/BF02296272
- Kim HY. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restor Dent Endod.* févr 2013;38(1):52-4. doi:10.5395/rde.2013.38.1.52 PubMed PMID: 23495371; PubMed Central PMCID: PMC3591587.
- Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ.* 27 juin 2011;2:53-5. doi:10.5116/ijme.4dfb.8dfd PubMed PMID: 28029643; PubMed Central PMCID: PMC4205511.
- Tavakol M, Brennan RL. Medical education assessment: a brief overview of concepts in generalizability theory. *Int J Med Educ.* 11 sept 2013;4:221-2. doi:10.5116/ijme.5278.a850 PubMed PMID: 28035055; PubMed Central PMCID: PMC4205529.
- Sadeghi P, Pourabbas A, Dehghani G, Katebi K. Quantitative and qualitative item analysis of exams of basic medical sciences departments of Tabriz University of Medical Sciences in 2023. *BMC Med Educ.* 1 juill 2025;25(1):937. doi:10.1186/s12909-025-07539-3 PubMed PMID: 40597222; PubMed Central PMCID: PMC12211469.
- Hermi A, Achour W. Item analysis of examinations in the Faculty of Medicine of Tunis. *Tunis Med.* avr 2016;94(4):247-52. PubMed PMID: 27704506.
- Ayub R, Kiran F, Shabnam N, Rauf A, Fatima F. Effects of a long term faculty development program on improvement in quality of MCQs: an impact evaluation study. *BMC Med Educ.* 15 avr 2025;25:541. doi:10.1186/s12909-025-07081-2 PubMed PMID: 40234907; PubMed Central PMCID: PMC12001514.
- Owolabi LF, Adamu B, Taura MG, Isa AI, Jibo AM, Abdul-Razek R, et al. Impact of a longitudinal faculty development program on the quality of multiple-choice question item writing in medical education. *Ann Afr Med.* 2021;20(1):46-51. doi:10.4103/aam.aam_14_20 PubMed PMID: 33727512; PubMed Central PMCID: PMC8102895.
- Abdulghani HM, Ahmad F, Irshad M, Khalil MS, Al-Shaikh GK, Syed S, et al. Faculty development programs improve the quality of Multiple Choice Questions items' writing. *Sci Rep.* 1 avr 2015;5:9556. doi:10.1038/srep09556 PubMed PMID: 25828516; PubMed Central PMCID: PMC4381327.
- Gupta P, Meena P, Khan AM, Malhotra RK, Singh T. Effect of Faculty Training on Quality of Multiple-Choice Questions. *Int J Appl Basic Med Res.* 2020;10(3):210-4. doi:10.4103/ijabmr.IJABMR_30_20 PubMed PMID: 33088746; PubMed Central PMCID: PMC7534721.
- Dellenges MA, Curtis DA. Will a Short Training Session Improve Multiple-Choice Item-Writing Quality by Dental School Faculty? A Pilot Study. *J Dent Educ.* août 2017;81(8):948-55. doi:10.21815/JDE.017.047 PubMed PMID: 28765439.
- Haladyna TM, Downing SM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Appl Meas Educ.* juill 2002;15(3):309-33. doi:10.1207/S15324818AME1503_5
- Pais J, Silva A, Guimarães B, Povo A, Coelho E, Silva-Pereira F, et al. Do item-writing flaws reduce examinations psychometric quality? *BMC Res Notes.* 11 août 2016;9:399. doi:10.1186/s13104-016-2202-4 PubMed PMID: 27516160; PubMed Central PMCID: PMC4982015.
- Shahat KA. Item Analysis of Multiple-Choice Question (MCQ)-Based Exam Efficiency Among Postgraduate Pediatric Medical Students: An Observational, Cross-Sectional Study From Saudi Arabia. *Cureus.* sept 2024;16(9):e69151. doi:10.7759/cureus.69151 PubMed PMID: 39262935; PubMed Central PMCID: PMC11388033.
- Bhattacharjee 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 Off Publ Indian Assoc Prev Soc Med.* 2022;47(1):92-5. doi:10.4103/ijcm.ijcm_1156_21 PubMed PMID: 35368481; PubMed Central PMCID: PMC8971860.
- Rezigalla AA, Eleragi AMESA, Elhussein AB, Alfaifi J, ALGhamdi MA, Al Ameer AY, et al. Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Med Educ.* 24 avr 2024;24(1):445. doi:10.1186/s12909-024-05433-y PubMed PMID: 38658912; PubMed Central PMCID: PMC11040895.
- Abdulghani HM, Irshad M, Haque S, Ahmad T, Sattar K, Khalil MS. Effectiveness of longitudinal faculty development programs on

- MCQs items writing skills: A follow-up study. PLOS ONE. 10 oct 2017;12(10):e0185895. doi:10.1371/journal.pone.0185895
24. Abozaid H, Park YS, Tekian A. Peer review improves psychometric characteristics of multiple choice questions. Med Teach. avr 2017;39(sup1):S50-4. doi:10.1080/0142159X.2016.1254743 PubMed PMID: 28103718.
 25. Karthikeyan S, O'Connor E, Hu W. Barriers and facilitators to writing quality items for medical school assessments - a scoping review. BMC Med Educ. 2 mai 2019;19(1):123. doi:10.1186/s12909-019-1544-8 PubMed PMID: 31046744; PubMed Central PMCID: PMC6498649.
 26. Ali K, Zahra D. Ten tips for effective use and quality assurance of multiple-choice questions in knowledge-based assessments. Eur J Dent Educ Off J Assoc Dent Educ Eur. mai 2024;28(2):655-62. doi:10.1111/eje.12992 PubMed PMID: 38282273.
 27. Keskes S, Ghazali H, Chermiti I, Ben Turkia H, Ben Salah N, Boujelbene N, et al. Analyse docimologique des grilles critériées dans une épreuve d'examen clinique objectif structuré en médecine d'urgence : évaluation de la validité et de la fiabilité. Tunis Med. 2026;104(2):281-287. doi:10.62438/tunismed.v104i02.6440.