

Analyse docimologique des grilles critériées dans une épreuve d'examen clinique objectif structuré en médecine d'urgence : Evaluation de la validité et de la fiabilité

Docimological analysis of criterion-referenced checklists in Objective Structured Clinical Examinations (OSCE) in acute medicine: Evaluation of validity and reliability

Syrine Keskes, Hanene Ghazali, Ines Chermiti, Hela Ben Turkia, Naoual Ben Salah, Nadia Boujelbene, Sami Souissi

Regional hospital of Ben Arous, Medical faculty of Tunis, Tunis El Manar university

RÉSUMÉ

Introduction : L'Examen Clinique Objectif et Structuré (ECOS) est un outil essentiel pour évaluer les compétences cliniques. Il est utilisé depuis 1998 à la Faculté de Médecine de Tunis, mais peu d'études ont analysé sa qualité docimologique, notamment en médecine d'urgence. Cette étude vise à évaluer la validité, la fiabilité et l'équité de l'ECOS, en mettant l'accent sur l'apport des grilles critériées.

Méthodes : Une étude prospective descriptive a été conduite en 2024–2025 auprès de 60 étudiants de DCEM3. L'épreuve comportait 12 stations couvrant savoir, savoir-faire et savoir-être, notées à l'aide de grilles critériées standardisées. Les analyses ont porté sur le taux de réussite, les indices de difficulté et de discrimination, le coefficient alpha de Cronbach et les paramètres de distribution.

Résultats : Soixante copies ont été analysées. Le taux de réussite global était de 96 %, avec une moyenne de $12,50 \pm 1,53/20$. L'indice de difficulté moyen (0,58) indiquait une difficulté acceptable et le coefficient alpha de Cronbach (0,85) reflétait une bonne homogénéité interne. L'indice de discrimination moyen était faible (0,16), mais les stations S11 et S12 (techniques et communication) présentaient un pouvoir discriminant satisfaisant, contrairement aux stations cognitives S1, S6 et S8. Les items de difficulté moyenne offraient la meilleure discrimination ($D \approx 0,21$). La corrélation difficulté–discrimination était quasi nulle ($r = -0,0316$).

Conclusions : L'ECOS en médecine d'urgence répond globalement aux recommandations docimologiques, avec difficulté équilibrée, bonne fiabilité et standardisation grâce aux grilles critériées. Certaines stations et items nécessitent une révision pour renforcer le pouvoir discriminant et améliorer la différenciation entre étudiants.

Mots clés : Docimologie, Pédagogie, Compétence

ABSTRACT

Introduction: The Objective Structured Clinical Examination (OSCE) is an essential tool for assessing clinical competencies. It has been used at the Faculty of Medicine of Tunis since 1998, yet few studies have analyzed its psychometric quality, particularly in emergency medicine. This study aims to evaluate the validity, reliability, and fairness of the OSCE, with a focus on the contribution of checklist-based evaluation grids.

Methods: A prospective descriptive study was conducted in 2024–2025 among 60 third-year medical students (DCEM3). The examination included 12 stations covering knowledge, skills, and professional behavior, scored using standardized checklist-based grids. Analyses focused on success rates, difficulty and discrimination indices, Cronbach's alpha, and score distribution parameters.

Results: Sixty student scripts were analyzed. The overall pass rate was 96 %, with a mean score of $12.50 \pm 1.53/20$. The average difficulty index (0.58) indicated acceptable difficulty, and Cronbach's alpha (0.85) reflected good internal consistency. The mean discrimination index was low (0.16), but stations S11 and S12 (technical and communication skills) showed satisfactory discriminative ability, unlike the cognitive stations S1, S6, and S8. Items of moderate difficulty offered the best discrimination ($D \approx 0.21$). The correlation between difficulty and discrimination was negligible ($r = -0.0316$).

Conclusion: The OSCE in emergency medicine generally meets psychometric recommendations, showing balanced difficulty, good reliability, and standardization through checklist-based grids. Certain stations and items require revision to enhance discriminative power and improve differentiation among students.

Key words: Docimology, Pedagogy, Competence

Correspondance

Syrine Keskes

Regional hospital of Ben Arous, Medical faculty of Tunis, Tunis El Manar university

Email: Syrine.keskes@fmt.utm.tn

INTRODUCTION

La docimologie, ou science de l'évaluation, occupe une place essentielle en pédagogie médicale. Elle permet d'analyser les processus par lesquels les compétences des apprenants sont mesurées, en mettant en évidence les biais et erreurs susceptibles d'altérer la validité et la fiabilité d'une épreuve (1). Dans un domaine où la formation de praticiens compétents est une exigence majeure, la docimologie contribue à garantir que les méthodes d'évaluation reflètent fidèlement les compétences cliniques attendues (2).

L'Examen Clinique Objectif et Structuré (ECOS), introduit par Harden et al.(3) en 1975, constitue une méthode d'évaluation standardisée des compétences cliniques. Contrairement aux examens oraux traditionnels, l'ECOS offre une évaluation plus objective et reproductible.

À la Faculté de Médecine de Tunis (FMT), l'ECOS a été introduit en 1998, sous la forme de l'évaluation clinique objective en stations multiples (ECOSM). Toutefois, peu d'études se sont attachées à analyser la qualité docimologique de ces épreuves à Tunis (5,6), notamment quant à l'apport des grilles critériées.

L'objectif de ce travail était de réaliser une analyse docimologique d'une épreuve d'ECOSM en médecine d'urgence destinée aux étudiants de troisième année du deuxième cycle des études médicales (DCEM3), et d'évaluer la contribution des grilles critériées dans l'amélioration de la validité, de la fiabilité et de l'équité de l'évaluation.

METHODES

Il s'agissait d'une étude prospective descriptive, menée durant l'année universitaire 2024–2025 à la Faculté de Médecine de Tunis. Nous avons inclus les étudiants de DCEM3 ayant effectué un stage de cinq semaines en médecine d'urgence au premier semestre et ayant passé l'épreuve d'ECOS correspondante. Les étudiants qui se sont absentés le jour de l'épreuve n'ont pas été inclus. Les étudiants dont les notes n'avaient pas été saisies selon le masque standard ont été exclus.

La médecine d'urgence fait partie du certificat de médecine aiguë enseignée à la 3ème année du 2ème cycle des études médicales de la Faculté de Médecine de Tunis. L'enseignement comporte un volet pratique obligatoire, reposant sur des stages cliniques (urgences, réanimation, réa-anesthésie).

Les étudiants de DCEM3 doivent ainsi effectuer un stage de cinq semaines dans un service des urgences, au cours duquel il doivent atteindre les objectifs pédagogiques détaillés dans le carnet de stage fourni par la faculté. Une évaluation périodique des activités dans le stage est complétée par une épreuve d'ECOS qui est une épreuve sanctionnelle avec une note éliminatoire de 8/10.

Description de l'épreuve ECOS

Choix des objectifs pédagogiques

Les thématiques des stations ont été élaborées à partir

des objectifs pédagogiques du carnet de stage de DCEM3 en médecine d'urgence couvrant trois domaines : savoir (n=10), savoir-faire (n=1), savoir-être (n=1).

Rédaction et organisation des stations

Un comité d'ECOS au sein de la section de médecine d'urgence était responsable de la finalisation de l'épreuve ECOS. Chaque encadrant référent a été sollicité pour produire des stations selon des thèmes bien définis. Une table de spécification a été élaborée afin de s'assurer de la couverture équilibrée des différentes dimensions de la compétence clinique (savoir, savoir-faire, savoir-être) dans l'ensemble des stations.

L'épreuve comprenait douze stations. La majorité (dix stations sur douze) étaient des stations écrites centrées sur les connaissances, principalement autour de l'interprétation d'examens complémentaires, du diagnostic et de la proposition de traitements.

Les stations comprenaient : des stations avec patient simulé (compétences relationnelles et cliniques), des stations avec mannequin (gestes techniques), des stations écrites (raisonnement clinique, interprétation d'électrocardiogrammes, de résultats biologiques ou radiologiques).

Deux stations pratiques avec observateur évaluaient le savoir-faire (la mise en place d'une CPAP) et le savoir-être (interrogatoire et communication avec le patient).

La durée a été standardisée à cinq minutes, afin d'offrir à chaque étudiant le même temps de réflexion et d'exécution.

Chaque station était notée sur dix points, décomposés en items précis, observables et mesurables. Les critères les plus essentiels pour la compétence visée ont bénéficié d'une pondération plus importante, allant jusqu'à deux points.

Constitution et vérification de la base de données

Les notes ont été saisies par les encadrants sur des fichiers Excel comportant un masque standardisé. Ces fichiers ont été fusionnés dans une base unique. Les notes par station étaient additionnées pour un total global sur 120, converti ensuite sur 20.

Une étape de vérification a été réalisée pour contrôler : la présence de toutes les copies, l'absence de doublons, la conformité des notes au barème, l'absence de cellules vides, la cohérence des totaux par station et de la note finale.

Analyse docimologique

Nous avons calculé les paramètres docimologiques permettant une analyse globale, une analyse par station et par item.

L'analyse globale de l'épreuve a porté sur la performance générale des étudiants, à travers le nombre ayant atteint la moyenne, le taux de réussite, ainsi que les notes minimale, maximale et la moyenne générale.

L'analyse par station a permis d'identifier, pour chacune d'elles, la moyenne obtenue ainsi que les valeurs extrêmes des notes.

Au niveau des items, l'évaluation a pris en compte la proportion de réponses nulles ou maximales. La validité des questions a été explorée à travers l'indice de difficulté et l'indice de discrimination.

La fiabilité interne de l'épreuve a été évaluée à l'aide du coefficient alpha de Cronbach, indicateur du degré d'homogénéité des items.

Enfin, l'analyse de la distribution des scores a reposé sur l'étude de l'asymétrie et de l'aplatissement de la courbe.

→ Indice de difficulté (p)

Cet indice mesure la validité d'une épreuve en exprimant le pourcentage de réussite à une question. Il définit la difficulté de la question. Il est représenté par le symbole p . L'indice de difficulté peut prendre des valeurs allant de 0 à 1.

Interprétation de l'indice de difficulté :

- $\leq 0,3$: la question est difficile
- Entre 0,3 et 0,7 : la question est à difficulté acceptable
- $\geq 0,7$: la question est facile

→ Indice de discrimination (ID)

Cet indice permet de mesurer la validité d'une épreuve. Il correspond à la capacité de l'item à départager les « forts » des « faibles ». Il est souvent représenté par le symbole (ID). Le calcul de l'indice de discrimination passe par deux étapes:

- Classer les étudiants selon la moyenne générale, et par ordre de mérite décroissant :
 - Le groupe « fort » : les 30% premiers.
 - Le groupe « faible » : les 30% derniers.
 - Le groupe « moyen » : les 40 % restants.
- Calculer la différence entre l'indice de difficulté d'un item pour le groupe dit « fort » ($p+$) et l'indice de difficulté pour le groupe dit « faible » ($p-$).

L'indice de discrimination (ID) sera égal à :

$$D = (p+) - (p-)$$

L'indice de discrimination varie de (-1) à 1.

Il est interprété selon Ebel (1965) (7):

- $\geq 0,4$: Item qui discrimine très bien
- 0,3-0,39 : Item qui discrimine bien
- 0,2-0,29 : Item qui discrimine peu
- 0,1-0,19 : Item-limite, mauvaise discrimination
- $< 0,1$: Item sans utilité réelle pour l'examen

Ont été considérés comme satisfaisants les items avec ID $\geq 0,30$.

→ Coefficient alpha de Cronbach

IL évalue l'homogénéité interne d'une épreuve.

Il varie de 0 à 1. Dans la pratique, on considère généralement que l'homogénéité de l'instrument est satisfaisante lorsque la valeur du coefficient est au moins égale à 0,8 [8].

L'interprétation des valeurs du coefficient de Cronbach selon Goerge et Mallery [9] :

- $\geq 0,90$: Excellente homogénéité interne
- Entre 0,80 et 0,90 : Homogénéité interne bonne ou satisfaisante
- Entre 0,70 et 0,79 : Homogénéité interne acceptable.
- $< 0,7$: homogénéité interne non acceptable

→ La courbe de distribution des scores

Dans notre étude, le chiffrier « AnItem.xls » utilisé nous a permis le traçage de la courbe de distribution des scores. L'analyse de la distribution des scores repose principalement sur deux indicateurs essentiels : le coefficient d'asymétrie et le coefficient d'aplatissement (10).

◊ Coefficient d'asymétrie

Le coefficient d'asymétrie a été calculé à l'aide de la fonction Excel. Cet indicateur permet d'apprécier la difficulté globale de l'épreuve :

- Asymétrie nulle : la distribution est symétrique et suit une loi normale (mode = moyenne = médiane). Les valeurs sont réparties de manière équilibrée de part et d'autre du centre.
- Asymétrie positive : la courbe est étalée vers la droite (mode < médiane < moyenne). Cela correspond à un effet plancher, traduisant un test globalement difficile.
- Asymétrie négative : la courbe est étalée vers la gauche (moyenne < médiane < mode). Cela correspond à un effet plafond, traduisant un test globalement facile.

◊ Coefficient d'aplatissement

Le coefficient d'aplatissement a été obtenu à l'aide de la fonction Excel KURTOSIS.

Il évalue le degré de concentration ou de dispersion des valeurs autour de la moyenne et indique si la discrimination se fait plutôt au centre de la distribution ou dans ses extrémités.

- Aplatissement négatif (distribution platykurtique): distribution basse et aplatie, avec un pic arrondi autour de la moyenne et des extrémités resserrées. La différenciation entre individus est alors plus difficile.
- Aplatissement positif (distribution leptokurtique) : distribution haute et pointue, avec un pic marqué au centre et des extrémités plus longues. La discrimination est difficile autour de la moyenne.
- Aplatissement nul (distribution mésokurtique) : situation intermédiaire entre les deux formes précédentes.

→ Corrélation difficulté discrimination

Elle a été calculée en utilisant le coefficient de corrélation de Pearson, applicable pour étudier la corrélation entre deux variables numériques à variation linéaire.

Analyse statistique

Les données ont été saisies à l'aide du logiciel Excel version 365 et analysées en utilisant le logiciel SPSS version 25.0. Les variables quantitatives ont été exprimées en moyenne $\pm$ écart type.

Le test t de student a été utilisé pour la comparaison de deux moyennes sur échantillons indépendants.

Considérations éthiques

Tous les étudiants ont donné leur consentement oral pour la participation à l'étude après avoir été informés du sujet et des objectifs du travail.

RÉSULTATS

Au total, 60 étudiants issus de huit services universitaires du Grand Tunis ont été inclus : Centre d'Aide Médicale Urgente (n=8), hôpital la Rabta (n=9), hôpital Habib Thameur (n=8), hôpital Charles Nicolle (n=8), hôpital régional de Ben Arous (n=7), hôpital Mahmoud Matri Ariana (n=7), hôpital des forces de sécurité intérieures (n=7), hôpital Mongi Slim (n=6). Sept étudiants ont été exclus pour données manquantes.

L'analyse a porté sur 60 copies et 138 items par copie.

L'analyse globale a mis en évidence un taux de réussite de 95 %, 51 ont obtenu une note $\geq 10/20$, avec une moyenne générale de $12,50 \pm 1,53/20$. Les résultats extrêmes variaient entre 9,08 et 14,42.

L'indice de difficulté global était de 0,58. L'indice de discrimination moyen était à 0,16. La distribution des scores présentait un coefficient d'aplatissement négatif (-1,089), traduisant une distribution platykurtique. Le coefficient d'asymétrie (-0,076), proche de zéro, montrait une répartition quasi symétrique des notes. L'histogramme accompagné d'une courbe de densité lissée et d'une courbe normale théorique confirme cette relative homogénéité (Figure 1). L'alpha de Cronbach était à 0,85.

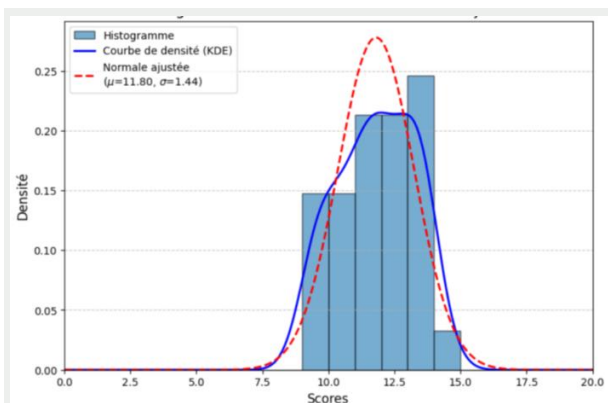


Figure 1. Courbe de la distribution des notes/20

Analyse par station

La moyenne des notes des stations écrites (stations de 1 à 10) et celles pratiques avec observateur étaient respectivement de $5,59 \pm 0,75$ et $7,77 \pm 1,55$ avec une différence significative ($p < 0,001$).

Trois stations (S2, S11, S12) se distinguaient par des scores particulièrement élevés. La majorité des autres stations oscillaient entre 5 et 6/10, illustrant une difficulté globalement modérée.

Concernant l'indice de difficulté par station, aucune station n'avait atteint les seuils extrêmes de 0 ou 1. Neuf stations étaient d'un niveau acceptable et trois étaient jugées faciles.

L'indice de discrimination par station était faible : la plupart se situaient entre 0,1 et 0,29. Seule la station 5 (station écrite) a démontré une discrimination satisfaisante. Quatre stations affichaient même un indice très faible ou négatif (figure 2).

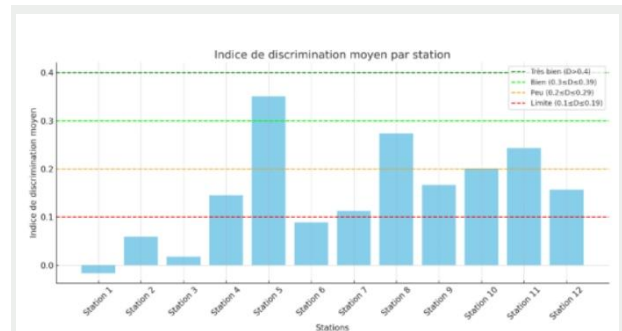


Figure 2. Répartition des stations selon leur indice de discrimination moyen illisible

L'étude des coefficients d'asymétrie montrait que la majorité des stations présentaient une asymétrie négative, parfois marquée (jusqu'à -1,36 pour S11), traduisant une tendance à la concentration des scores élevés. Seule la station 7 a affiché une asymétrie positive notable (+0,29). Plusieurs stations, comme S1, S2 ou S9 étaient proches de la symétrie parfaite.

La majorité des stations avaient un kurtosis négatif ou légèrement positif, situé entre -1 et +2.

Quelques stations (S8; S11) avaient un kurtosis plus élevé, traduisant une distribution plus concentrée avec quelques valeurs extrêmes.

Analyse par item

L'analyse par item a montré que plus que la moitié (56,45 %) des items présentaient au moins 30 % de notes nulles. Les stations 5, 6 et 9 se distinguaient par des taux particulièrement élevés (jusqu'à 80 % pour S6).

Cependant, presque un quart des items ont 80 % ou plus de notes maximales. Trois stations (S1, S5, S10) n'en comptaient aucun, tandis que S12 en présentait 58 %.

La plupart des items se situaient dans une zone de difficulté acceptable, assurant un équilibre général. Toutefois, les stations S1, S7 et S10 avaient un pourcentage important d'items difficiles ($< 0,3$). Les stations S2, S3, S5 et S8 ont montré une répartition harmonieuse entre items faciles et modérés (figure 3).

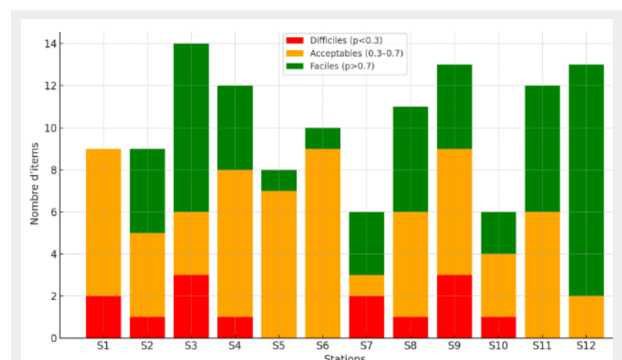


Figure 3. Répartition des items selon leur indice de difficulté

La majorité des stations avaient beaucoup d'items à faible pouvoir discriminant ($< 0,1$ ou $0,1-0,19$).

Les items « très bien discriminants » ($> 0,4$) restaient rares et concentrés sur quelques stations.

Les Stations 5, 8 et 11 ressortaient avec plusieurs items très performants (3 ou plus) (figure4).

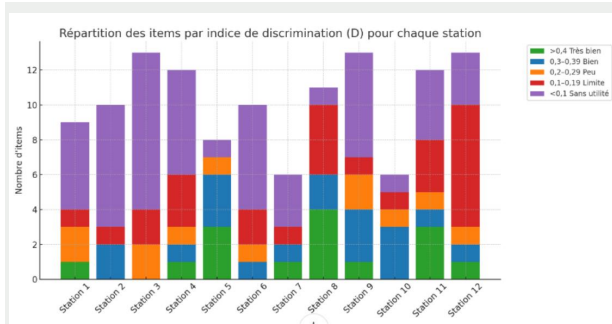


Figure 4. Répartition des items selon leur indice de discrimination

Corrélation entre difficulté et discrimination

L'étude de la relation entre ces deux indices a révélé un coefficient de corrélation de $-0,0316$, traduisant une absence quasi totale de lien : la difficulté des items n'influence pas leur capacité à discriminer. Le graphique de corrélation visualise cette relation faible entre difficulté et discrimination pour l'ensemble des items analysés (figure 5).

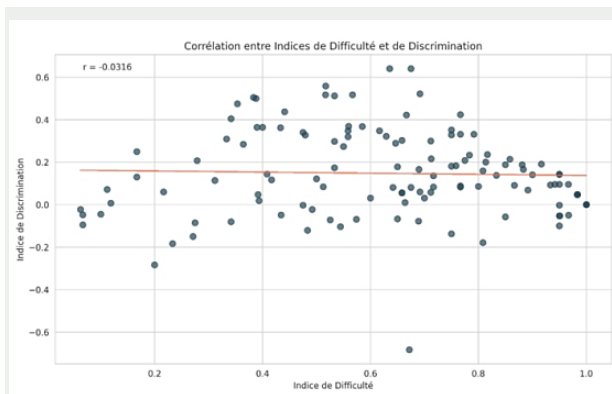


Figure 5. Graphique de corrélation entre indices de difficulté et de discrimination

Néanmoins, les items de difficulté moyenne présentaient une meilleure capacité de discrimination. La majorité des items se situaient dans la catégorie de difficulté moyenne. Le coefficient de discrimination moyen entre les différentes stations selon le niveau de difficulté est représenté dans le tableau 1.

Tableau 1. Coefficient de discrimination des différentes stations selon le niveau de difficulté

Niveau de difficulté	Coefficient de discrimination moyen	Nombre d'items
Facile	0,117	50
Moyenne	0,211	59
Difficile	-0,013	14

DISCUSSION

Dans notre étude concernant 60 étudiants, le taux de réussite était élevé de 96 %, avec une moyenne générale de $12,50 \pm 1,53/20$ et des notes comprises entre 9,08 et 14,42.

L'indice de difficulté global était à 0,58 indiquant un niveau adapté et reflétant une épreuve ni trop facile ni trop difficile.

L'indice de discrimination global était à 0,16 montrant une capacité limitée de l'épreuve à différencier les étudiants selon leur niveau de compétence.

L'étude de la distribution des scores confirme ces observations. La répartition est quasi symétrique, platykurtique et globalement ajustée à une loi normale, traduisant l'absence d'effets de seuil et une homogénéité des performances. La concentration des notes autour de 12–13/20, sans valeurs extrêmes, atteste de la faisabilité de l'épreuve et de son adéquation avec le niveau global des étudiants. Toutefois, la dispersion réduite des scores limite la différenciation entre les étudiants les plus performants et ceux en difficulté.

Enfin, la fiabilité interne, attestée par un coefficient alpha de Cronbach de 0,85 a confirmé une bonne cohérence des items et une solidité de l'évaluation sur le plan méthodologique.

Discussion des résultats

Le nombre de stations est un déterminant majeur de la fiabilité d'un ECOS. Dans notre étude, le nombre total de stations était fixé à 12, ce qui se situe au seuil inférieur recommandé par la littérature. Plusieurs travaux montrent que l'augmentation du nombre de stations contribue davantage à améliorer la fiabilité que la prolongation de la durée de chaque station [12,14]. Ainsi, un ECOS comportant un nombre limité de stations peut manquer de représentativité vis-à-vis de l'ensemble des compétences attendues.

Le taux élevé de réussite observé dans notre étude rejoint les données internationales. En effet, les ECOS nationaux français de 2025 ont montré moins de 1 % d'échec [9].

Une étude britannique a rapporté que les taux de réussite aux ECOS variaient entre 85 et 100%, en fonction des critères de validation de l'épreuve [10].

Dans une étude tunisienne concernant une épreuve ECOS en nutrition, le taux de réussite était de 100% [5].

Le taux élevé de réussite observé peut s'expliquer par plusieurs facteurs liés à la qualité de la préparation et de l'organisation de l'ECOS. L'utilisation systématique d'une table de spécification permet d'assurer une répartition équilibrée des stations et des compétences évaluées, en adéquation avec les objectifs pédagogiques visés.

L'indice de difficulté moyen de 0,58 se situait dans l'intervalle optimal recommandé (0,3–0,7), garantissant une évaluation ni trop facile ni trop exigeante [11].

L'indice de discrimination moyen de 0,16, bien que légèrement inférieur aux valeurs recommandées ($ID > 0,2$) reste comparable à plusieurs études ayant analysé des ECOS dans des contextes similaires [12,13].

La distribution des scores quasi symétrique et platykurtique observée dans notre étude rejoint également les résultats publiés dans la littérature, où de nombreux auteurs rapportent des distributions aplaties lors d'évaluations standardisées en médecine [12].

La dispersion relativement réduite des scores limite partiellement la différenciation entre les étudiants les

plus performants et ceux en difficulté, ce qui suggère un test davantage centré sur la validation des acquis de base que sur la sélection des meilleurs candidats.

L'analyse des résultats des stations écrites et pratiques montre une différence significative entre les scores moyens des stations pratiques et les stations écrites. Ce phénomène est cohérent avec la littérature qui souligne que les stations pratiques évaluées directement par un observateur peuvent bénéficier d'une plus grande flexibilité dans l'attribution des points et d'une interaction favorisant la performance des étudiants [12].

Le recours à un nombre important de stations écrites dans un ECOS permet de diversifier l'évaluation et d'assurer une meilleure standardisation [3]. Toutefois, un excès de stations écrites risque de détourner l'examen de sa finalité initiale, centrée sur l'évaluation des compétences pratiques et relationnelles, en le rapprochant d'une épreuve théorique de type QCM [14]. Cette prédominance soulève également un problème d'authenticité, les étudiants pouvant mobiliser leurs connaissances sans démontrer leur performance clinique réelle [14]. La littérature recommande donc de maintenir un équilibre, avec une proportion de stations écrites limitée à environ un tiers, afin de garantir la validité et la représentativité de l'épreuve [12,15].

Etude de la fiabilité : coefficient Alpha de Cronbach

Le coefficient Alpha de Cronbach obtenu était de 0,85, ce qui témoigne d'une bonne homogénéité interne de l'épreuve.

En psychométrie, un alpha supérieur à 0,7 est généralement considéré comme acceptable, tandis qu'une valeur supérieure à 0,8 traduit une bonne cohérence interne entre les items ou stations d'un test [15,16].

Dans le domaine des ECOS, des valeurs comprises entre 0,7 et 0,8 sont généralement considérées comme acceptables [12,15]. En effet, il est reconnu que les ECOS évaluent une large gamme de compétences et ne sont pas des tests purement homogènes comme une QCM. Ce résultat suggère que le nombre et la diversité des stations étaient suffisants pour assurer une bonne fiabilité, tout en maintenant la validité du contenu.

En termes de difficulté, aucune station n'avait des valeurs extrêmes de 0 ou 1 et la majorité se situait dans l'intervalle recommandé de difficulté modérée (0,3–0,7). Ce constat correspond aux recommandations générales en évaluation, où un indice de difficulté moyen autour de 0,5–0,6 est jugé optimal [12,13].

La station 1 présente l'indice le plus bas traduisant une difficulté relativement accrue pour les étudiants. Il s'agit d'une question récemment introduite dans la banque des questions et qui propose une tâche complexe mieux adaptée à une évaluation écrite qu'à une épreuve pratique. À l'inverse, la station 12 jugée la plus facile est une station pratique fréquemment proposée dans les ECOS.

Concernant l'indice de discrimination, la plupart des stations présentent des valeurs faibles indiquant une capacité limitée à différencier les étudiants selon leur niveau réel. Des indices de discrimination faibles sont

fréquents dans les ECOS, notamment lorsque les stations sont trop faciles, trop standardisées, ou comportent peu d'items cotés.

Analyse par item

Indice de difficulté par item

L'analyse des indices de difficulté montre une répartition en faveur des items dans la zone intermédiaire (52,8%). La littérature recommande qu'environ 50–70 % des items soient de difficulté intermédiaire pour une bonne discrimination [11].

Mahjoubi et al.[5] dans leur analyse docimologique de l'ECOS en Nutrition à la Faculté de Médecine de Tunis ont également observé une répartition similaire. Plus que la moitié des questions (53%) étaient faciles et 47% des questions étaient de difficulté acceptable.

Indice de discrimination par item

L'examen des indices de discrimination des douze stations a mis en évidence une hétérogénéité importante dans la capacité des items à différencier les candidats selon leur niveau de compétence. La majorité d'items présentaient un pouvoir discriminant faible.

Selon Ware et Vik [17], au moins 60 % des items devraient présenter une bonne discrimination. Dans notre étude, 28 % des items étaient excellents ou bien discriminants et 65 % présentaient une discrimination faible ou nulle.

Dans le cadre de la pédagogie de la maîtrise, certains objectifs d'apprentissage doivent être acquis par la quasi-totalité des étudiants (80 à 90 %) avant d'aborder les objectifs suivants. Dans une telle approche, la fonction discriminative des items perd de sa pertinence : ceux-ci devraient en effet être réussis par la grande majorité des étudiants après l'enseignement [18].

L'analyse des coefficients de discrimination en fonction du niveau de difficulté des items a montré que les items de difficulté moyenne présentaient une meilleure capacité de discrimination ($D = 0,211$) que les items faciles ($D = 0,117$) ou difficiles ($D = -0,013$).

Ce constat est cohérent avec la littérature en psychométrie, qui souligne que les items modérément difficiles sont généralement les plus informatifs pour distinguer les performances des candidats [3,11].

En effet, les items très faciles sont souvent réussis par la majorité des candidats, ce qui limite leur pouvoir discriminant, tandis que les items très difficiles sont rarement réussis, réduisant également leur capacité à différencier les niveaux de compétence.

Dans notre étude, la majorité des items se situe dans la catégorie de difficulté moyenne (59 items sur 123), ce qui favorise une discrimination plus efficace entre les étudiants et contribue à la validité de l'épreuve.

Importance des grilles critériées dans les ECOS

Les grilles critériées représentent un outil essentiel pour garantir la validité, la fiabilité et la faisabilité des ECOS. Elles standardisent l'évaluation des compétences spécifiques, facilitent l'appropriation par les enseignants, même novices, et permettent le recours à des observateurs non experts, ce qui est particulièrement utile en contexte de ressources limitées[19].

Comparées aux grilles globales, elles offrent une meilleure fidélité inter-juges, à condition d'être soigneusement construites et centrées sur les points clés. Elles sont particulièrement adaptées aux compétences procédurales, tandis que pour les dimensions relationnelles ou organisationnelles, l'usage d'échelles qualitatives est recommandé [20].

CONCLUSIONS

L'évaluation de l'évaluation est indispensable pour garantir la validité, la fiabilité et l'équité des ECOS. Elle permet d'identifier les forces et limites des épreuves, d'optimiser les outils (grilles, formation des examinateurs, standardisation) et d'assurer une certification juste et crédible.

L'amélioration de l'épreuve passe par la formation des enseignants, l'optimisation des grilles critériées, l'exploitation régulière des indicateurs docimologiques et le retour systématique des résultats aux évaluateurs.

RÉFÉRENCES

1. LEclercq D, Nicaise J, Demeuse M. Docimologie critique: des difficultés de noter des copies et d'attribuer des notes aux élèves. In: Demeuse M, éditeur. Introduction aux théories et aux méthodes de la mesure en sciences psychologiques et en sciences de l'éducation. Liège: Les éditions de l'Université de Liège; 2004. p. 273-92.
2. De Ketele JM. L'évaluation conjuguée en paradigmes. Revue française de pédagogie. 1993;103:59-80.
3. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;1(5955):447-51.
4. Chisnall B, Vince T, Hall S, Tribe R. Evaluation of outcomes of a formative objective structured clinical examination for second-year UK medical students. Int J Med Educ. 2015;6:76-83.
5. Mahjoub F, Mizouri R, Amor NB, Ounaissa K, Achour W, Jamoussi H. L'évaluation clinique objective en stations multiples en nutrition: analyse docimologique. Tunis Med. 2020;98:548-54.
6. Limaïem F, Farah F, Belkacem O, Bouraoui S, Achour W. Objective multi-station clinical evaluation in pathological anatomy: critical analysis and perception. Tunis Med. 2020;98:674-81.
7. Martin J. Aux origines de la « science des examens » (1920-1940). Hist Educ. 2002;(94):177-99.
8. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297-334.
9. George S, Haque MS, Oyebo F. Standard setting: comparison of two methods. BMC Med Educ. 2006;6:46.
10. Carricano M, Poujol F. Analyse de données avec SPSS. Pearson. Paris;2010.p.33-8
11. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-7.
12. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: a review of metrics. AMEE guide no. 49. Med Teach. 2010;32:802-11.
13. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003;80(1):99-103.
14. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ. 2004;38(2):199-203.
15. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach. 2013;35:503-14.
16. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994. 786 p.
17. Ware J, Vik T. Quality assurance of medical multiple-choice questions by a simple computerized analysis. BMC Med Educ. 2009;9:40.
18. Haladyna TM, Rodriguez MC. Developing and validating test items. New York: Routledge; 2013. 454 p.
19. Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ. 2015;49(2):161-73.
20. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med. 1998;73(9):993-7.