

Introduction

La formation continue des médecins implique la lecture régulière de travaux scientifiques ; les médecins doivent, en effet, pouvoir mettre à jour leur documentation professionnelle, ce qui implique en particulier d'analyser et juger de façon autonome la valeur scientifique des articles publiés, ainsi que leur caractère applicable à d'autres situations.

L'enseignement de la lecture critique d'article (LCA) vise à compléter leur formation initiale dans ce domaine ; elle substitue partiellement l'acquisition d'un savoir (connaissances) par celle d'un savoir-faire (sélectionner soi-même des connaissances valides).

Cet enseignement, qui doit préparer les futurs médecins à leur formation continue, est différent de l'expertise scientifique requise en amont, pour sélectionner un manuscrit en vue d'une éventuelle publication scientifique¹ (processus éditorial).

Contexte

Une épreuve de LCA des étudiants est prévue dans le cadre de l'examen classant national (ECN) à partir de 2008², selon des modalités définies par le Centre national du concours de l'internat (CNCI). Cette nouvelle épreuve de trois heures, comptant pour 20 % des points, devrait concerner un article original, de portée générale, écrit en langue française ; elle devrait associer six à dix questions sur la méthodologie du travail, d'une part, et la rédaction d'un résumé de moins de 250 mots, d'autre part.

La plupart des facultés de médecine ont mis en place de façon empirique dès 2002 des formations à cette épreuve ; toutefois, ni les étudiants, ni les correcteurs ne sont entraînés à de telles épreuves nouvelles, alors qu'une telle expérience est nécessaire pour évaluer l'acquisition³ de cette habileté. L'organisation d'une épreuve de LCA dans un examen national à visée normative semble préoccupante dans un tel contexte, ce d'autant que la reproductibilité de la correction d'une épreuve de LCA n'est pas connue.

Le but de cette étude-pilote est d'apprécier la reproductibilité intercorrecteurs d'une correction d'épreuve de LCA organisée suivant les spécifications actuellement connues de cette épreuve, sans précaution méthodologique particulière, dans les conditions actuelles de formation des étudiants et de préparation des correcteurs, afin de repérer les difficultés potentielles de ce type d'épreuve.

Sujets et méthodes

Etudiants

À la faculté de médecine de Rouen, un enseignement des méthodes épidémiologiques est réalisé depuis 1998 en première année du deuxième cycle des études médicales (DCEM1) sous forme d'un autoapprentissage des principes de méthode, suivi de séances d'enseignement dirigé de 90 minutes, utilisant des articles originaux comme supports ; cet enseignement est validé par une épreuve terminale d'analyse d'article de 30 minutes, comportant 4 à 5 questions concernant un résumé d'article⁴. Les étudiants ont été sollicités par voie d'affichage et 59 volontaires ont passé l'épreuve en septembre 2002 : 25 et 34 étaient respectivement en deuxième et troisième années du DCEM, soit respectivement 25 et 31 % des inscrits des promotions concernées.

Ces étudiants ont été soumis à une épreuve dans les conditions prévues pour l'ECN, mais sans formation supplémentaire, en particulier pour la rédaction d'un résumé.

Dossier de LCA

Un article original a été choisi par un groupe local de cinq enseignants comme support à une épreuve de LCA : il s'agissait d'un essai randomisé bicentrique comparant deux modes de dépistage de l'hépatite C en médecine générale, selon que les patients avaient reçu ou non une information préalable⁵. Un dossier comportant l'article (sans le résumé ni la discussion), et dix questions (*Tableau 1*), y compris les réponses-type attendues (*cf. annexe I*) et un résumé type (*cf. annexe II*), a été construit⁶ ; il a fait l'objet d'un travail commun des enseignants au cours de 5 réunions. Ce dossier a été ensuite validé par un groupe d'experts de LCA auprès du CNCI au niveau national.

Epreuve

L'épreuve a été organisée dans les conditions d'un concours : surveillance, absence d'échanges entre les étudiants, temps limité à 3 heures, remise d'une copie rendue anonyme. Les étudiants avaient pour consigne : a) de répondre aux questions posées concernant l'article ; b) de rédiger un résumé selon les standards admis de la rédaction scientifique. La consigne précisait explicitement que le résumé devait comprendre des paragraphes respectivement dédiés aux objectifs, aux méthodes, aux résultats et à la conclusion.

Correction

Trois correcteurs (X, Y et Z), enseignants à la faculté, appartenant au groupe ayant conçu le dossier, ont corrigé chacun les 59 copies de façon indépendante, à l'aide d'un barème sur 100 points, répartis entre les dix questions et le résumé (*tableau 1*).

Analyse statistique

Les scores portés par chaque correcteur à chaque étudiant ont fait l'objet d'une double saisie. Les scores moyens des 3 correcteurs aux mêmes copies pour les questions, d'une part, et pour le résumé, d'autre part, ont été comparés par un test *t* apparié. Les coefficients de corrélation intraclasses⁷ (CCI) et leur intervalle de confiance à 95 % (IC 95 %) ont été calculés pour les trois couples de correcteurs pour chaque question, pour l'ensemble des questions, pour le résumé et enfin pour l'ensemble de l'épreuve. Les CCI ont été comparés par étude du recouvrement des IC 95 %. Le niveau de concordance a été qualifié selon les références admises. La moyenne des CCI par correcteur a été calculée. La relation entre la reproductibilité des corrections et le niveau des copies a été testée par un coefficient de corrélation entre la moyenne des CCI et celle des scores obtenus par les étudiants.

Résultats

Le scores moyen (\pm écart type) sur 20 donné par les 3 correcteurs à l'ensemble de l'épreuve aux 59 étudiants était de 8,7 (\pm 2,0) avec une dispersion modérée (minimum 4,5 ; maximum : 14,6). Le scores moyen était de 8,2 (\pm 2,2) pour les questions et de 9,9 (\pm 2,6) pour le résumé. L'écart moyen entre les scores du résumé et des questions de chaque étudiant était de 1,7 (\pm 2,6), écart supérieur à zéro ($p < 10^{-6}$).

Les CCI des questions, du résumé et de l'ensemble des épreuves figurent dans le *Tableau 2* pour les 3 paires de correcteurs. Pour l'ensemble de l'épreuve, les résultats variaient de 0,48 (IC 95 % : 0,26 – 0,65) pour la paire YZ à 0,72 (IC 95 % : 0,55 – 0,81) pour la paire XZ.

Sur les 48 CCI (16 items pour 3 paires), aucun n'était très bon, 8 étaient bons, 14 modérément bons, 16 médiocres et 10 mauvais selon les qualificatifs usuels.

La correction des questions était plus reproductible que celle des résumés pour 2 des 3 paires de correcteurs.

La reproductibilité (moyenne des CCI des trois paires de correcteurs) n'était pas significativement liée avec le score moyen obtenu par les étudiants ($r = -0,35$).

Discussion

Ce travail pilote comporte des limites notables : les étudiants étaient volontaires, issus d'une seule faculté et peu nombreux ; ils n'étaient concernés par l'ECN que 2 à 3 ans plus tard, délai permettant des apprentissages complémentaires ; leur habilité⁸, telle qu'elle a été évaluée dans ce travail, ne correspond pas à celle correspondant à la date prévue pour eux pour cet examen. L'absence d'enjeu pour eux modifiait les conditions de l'épreuve et peut-être leur comportement. L'analyse ne portait que sur un seul article, corrigé par des enseignants d'une seule faculté, la plupart n'ayant pas une expérience notable de corrections d'épreuves de LCA.

Ces étudiants n'avaient pas encore fait l'objet d'une préparation spécifique de la LCA en vue de l'ECN, mais ils avaient reçu une formation méthodologique en DCEM1, formation s'appuyant en bonne partie sur l'analyse d'articles scientifiques. Toutefois, cette situation correspond probablement à celle d'autres facultés en France actuellement ; elle est donc opportune pour juger des performances qui pourraient être associées à une telle épreuve, sans précautions particulières.

Dans ces conditions, ce groupe d'étudiants de 2^e cycle n'a pas obtenu la moyenne à cette épreuve de LCA respectant les conditions définies par le CNCI ; l'épreuve est donc apparue difficile. Le score moyen obtenu au résumé était paradoxalement meilleur que pour les questions, alors que les étudiants n'avaient pas été formés spécifiquement à cet exercice⁹.

L'épreuve est également apparue peu discriminante, d'après la dispersion des scores. En outre, sa correction a été difficile, comme le suggère la faible reproductibilité inter-juges. La reproductibilité du processus de correction pouvait être influencée par l'épreuve elle-même, par les items qui la composaient, par les étudiants ou par les évaluateurs. La valeur moyenne des CCI des questions et du résumé étaient proches, correspondant à un accord inter-juges modéré ; ce résultat médiocre n'était pas le fait de l'une ou l'autre des deux parties de l'épreuve. La reproductibilité de la correction variait fortement selon les items : elle était médiocre ou mauvaise pour plus de la moitié d'entre eux, mais surtout la dispersion des CCI était importante (extrêmes : - 0,07 à 0,90).

Certaines difficultés de correction imprévues ont été rencontrées pour certains items, mettant en difficulté les trois correcteurs (Q1c et conclusion du résumé) ; on peut s'interroger sur le libellé de telles questions, voire leur contenu réel.