Difficultés de correction d'une épreuve d'analyse critique d'article scientifique : une étude exploratoire

Christine LOUIS-SYLVESTRE [1], Claire FURHMAN [2], Bruno HOUSSET [2]*

Résumé ***Contexte :*** *L'apprentissage de la lecture critique d'article est évalué dans notre faculté par une épreuve comportant la réalisation d'un résumé d'article. Les résultats obtenus par les étudiants lors des premières éditions de cette épreuve n'étaient apparemment pas corrélés à leur niveau et n'étaient pas reproductibles d'une épreuve à l'autre. Ceci nous a fait suspecter des difficultés de correction de l'épreuve.* ***Méthode :*** *20 copies ont été tirées au sort et évaluées par 5 correcteurs à l'aide d'une même grille. Nous avons calculé, pour chaque copie, la moyenne des 5 scores obtenues et calculé un pourcentage de variation. Nous avons comparé les 5 correcteurs en calculant pour chacun d'eux la moyenne des 20 scores qu'ils avaient attribués. La concordance entre les différents examinateurs a été étudiée grâce au test de Kendall. Les analyses ont été réitérées avec une deuxième grille de correction plus détaillée. Les pourcentages de variation observés avec chacune des deux grilles ont été comparés par un test du X^2.* ***Résultats :*** *Les scores varient de 45 % (extrêmes 17 à 88 %) selon le correcteur avec la première grille, et de 32 % (10 à 60 %) avec la deuxième grille. Les correcteurs sont cependant concordants dans le classement des copies ($p < 0,001$). Le pourcentage de variation inter-correcteur des scores est significativement réduit avec la deuxième grille ($p < 0,01$).* ***Conclusion :*** *La réalisation d'un résumé à partir d'un article est une épreuve difficile à corriger. L'utilisation d'une grille de correction, même si elle est détaillée, laisse persister des variations entre correcteurs.*

Mot clés Etudes médicales ; lecture critique d'article ; évaluation ; correction ; résumé d'article.

Abstract ***Context:*** *Critical appraisal skills are partly evaluated in our university by asking the students to elaborate a summary of a scientific paper. The results obtained by the students for the first evaluations showed a lack of correlation with the level of the student and were not reproducible. This led us to suspect difficulties in correcting this examination.* ***Method:*** *Twenty summaries were taken at random and evaluated by 5 correctors with the same correction frame. We calculated the mean mark obtained by each summary and the percentage of variation. We compared the 5 correctors by calculating the mean mark they attributed to the 20 summaries. Kendall's test was used to study the concordance between all the correctors. The same analysis was reiterated with a second correction with a more detailed frame. Percentages of variation with each frame were compared with a chi square test.* ***Results:*** *The percentage of variation of the marks between correctors was 45% (range 17-88%) with the first frame, and 32% (10-60%) with the second. Correctors were however concordant for the rank of the copies ($p<0.001$). Percentage of variation of the note according to the corrector was significantly reduced with the second frame ($p<0.01$).* ***Conclusion:*** *Elaboration of a summary is a test that proves difficult to correct. With a frame, however detailed it is, variations between correctors do not disappear.*

Key words Medical under graduate curriculum; critical appraisal skills; assessment; paper abstract; rating.

Pédagogie Médicale 2005; 6: 138-146

1- Service de gynécologie - Centre Hospitalier Intercommunal - Créteil - France
2- Service de pneumologie - Centre Hospitalier Intercommunal - Créteil - France
** Vice doyen de la Faculté de médecine de Créteil*
Correspondance : Christine Louis-Sylvestre – Maternité Centre Hospitalier Intercommunal – 40, Avenue de Verdun 94000 - Créteil - France - Tel : 01 45 17 55 51 - mailto : christine.louissylvestre@chicreteil.fr

Introduction

La lecture régulière et critique de la littérature scientifique s'avère indispensable à la formation des médecins. Elle est cohérente avec le développement récent du concept d'*« evidence-based medecine »*, autrement dit de pratique médicale fondée sur les résultats d'études scientifiques. La pléthore et la qualité très inégale des études publiées imposent que chacun pratique l'analyse critique de ces articles. Peu d'études ont analysé les effets de cette pratique et aucune n'a pu démontrer clairement un impact positif sur la qualité des soins apportés au patient[1-5]. Néanmoins, la lecture de la littérature médicale est un des moyens dont disposent les praticiens pour se tenir au courant des évolutions et s'inscrit tout naturellement dans le processus de formation continue. L'apprentissage de la pratique de la lecture critique deviendra une obligation pendant les études de médecine en France et fera l'objet d'une épreuve au cours de l'examen national classant. La plupart des facultés de médecine ont donc organisé un enseignement de cette pratique et, en corollaire, un contrôle de l'acquisition de ces connaissances.

La formation à la lecture critique d'articles a débuté à la faculté de médecine de Créteil en 2002, pendant le deuxième cycle des études médicales. Cette séquence d'enseignement et d'apprentissage a fait l'objet d'un examen écrit ; celui-ci a eu lieu deux fois, en février et en juin. Il consistait en la lecture d'un article après laquelle il était demandé à l'étudiant de composer le résumé de l'article puis de répondre à des questions rédactionnelles. Une première analyse des scores attribués (résultats non publiés) a mis en évidence l'absence de corrélation entre les résultats obtenus par les étudiants dans les autres disciplines et les scores obtenus à l'épreuve de lecture critique d'article. Il n'y avait pas non plus de corrélation entre les scores de février et ceux de juin, ni entre le score obtenu au résumé et celui obtenu aux questions rédactionnelles.

Enfin, les scores des résumés couvraient une échelle plus large que celle des questions rédactionnelles. Nous avons pensé que toutes ces discordances apparentes pouvaient traduire, entre autres, une difficulté de correction de l'épreuve. En particulier, nous nous sommes demandés s'il n'y avait pas eu de grandes différences dans les modalités de correction des résumés suivant les correcteurs. Nous avons donc décidé d'analyser les résultats obtenus à l'épreuve de résumé lorsque les copies étaient cotées par des correcteurs différents utilisant une même grille de correction. Ce sont les résultats de cette étude que nous présentons ici.

Matériel et méthodes

L'épreuve de lecture critique d'article a concerné 92 étudiants en 2002, à deux reprises, en février puis en juin. A chaque fois, il leur était soumis un article dont le résumé et la discussion avaient été masqués. L'épreuve, effectuée en temps limité, comportait deux sections : les étudiants devaient faire un résumé de l'article puis répondre à 14 questions rédactionnelles. La technique du résumé leur avait été enseignée en cours d'année avec, en particulier, étude de la structure type (introduction, matériel et méthodes, résultats, conclusion). Nous nous sommes intéressés à l'épreuve de juin. L'article choisi rapportait les résultats d'une étude prospective randomisée concernant la prévention d'accidents cardio-vasculaires dans une population à haut risque. Etant donné le programme des cours dispensés dans l'année, l'article était compréhensible et analysable par tous les étudiants, quel que soit leur parcours en stage hospitalier. Parmi les 92 copies, 20 ont été tirées au sort et soumises à cinq correcteurs n'ayant pas participé à la correction officielle.

Pour l'étude que nous présentons ici, nous nous sommes intéressés uniquement au résumé élaboré par les étudiants. Chacune des 20 copies a été évaluée par cinq correcteurs différents auxquels il avait été demandé de corriger en se rapportant à la grille fournie en *annexe 1* . Cette grille a été bâtie à partir de la structure type d'un résumé, structure que les étudiants devaient respecter. Le score global attribué au résumé était décomposé par paragraphes correspondant à la structure type du résumé.

Une grille de correction efficace doit notamment faire en sorte que le score obtenu soit indépendant du correcteur ; en termes métrologiques, ce point rend compte de la fidélité inter-correcteur. Nous avons donc analysé la dispersion des cinq scores obtenus par chaque copie en calculant la moyenne (± écart type) sur ces cinq scores et en calculant un pourcentage de variation (différence entre les deux scores extrêmes divisée par la moyenne). Puis, nous avons classé ces pourcentages de variation en 4 classes (< 20 %, 20 à 40 %, 40 à 60 %, > 60 %).

Nous avons ensuite comparé les manières de coter des correcteurs. Pour ceci, nous avons calculé la moyenne ± écart type des scores attribués par chaque correcteur. Nous avons aussi regardé combien de fois il était arrivé à chaque correcteur de donner un score sortant de la moyenne ± écart type des cinq scores obtenus par la copie.

Pour vérifier qu'il y avait une concordance entre les différents examinateurs, nous avons utilisé le test statistique de concordance de Kendall. Ce test vérifie que les correcteurs,

tout en ne donnant pas forcément des scores identiques, classent cependant les copies dans le même ordre. Pour effectuer ce test, un rang a été affecté à chaque copie selon son score.
Au vu des résultats observés avec la première grille, nous avons décidé de réitérer l'étude avec une deuxième grille, cette fois-ci plus détaillée. Celle-ci a été élaborée en tenant compte des difficultés de cotation qu'avaient pu signaler les différents correcteurs. Elle figure en *annexe 2*. Le même test de Kendall a été utilisé pour l'analyse de ces résultats. Pour comparer l'efficacité des deux grilles, nous avons analysé la répartition des copies entre les quatre classes de coefficient de variation (< 20 %, 20 à 40 %, 40 à 60 %, > 60 %). Cette comparaison a été effectuée grâce à un test du X^2. Les résultats ont été jugés statistiquement significatifs pour $p < 0,05$.

Résultats

Résultats concernant les corrections utilisant la première grille

Pour une copie donnée, le pourcentage de variation de la note selon le correcteur est de 45 % en moyenne (extrêmes 17 à 88 %). Ces variations sont observées autant sur les six moins bonnes copies que sur les six meilleures. Le chapitre sur lequel il y a le moins de différence de cotation entre les correcteurs est le chapitre « Matériel et Méthodes ». Le chapitre qui génère la plus grande variation de cotation est le chapitre « Introduction».
La moyenne des scores donnés à l'ensemble des copies diffère selon le correcteur ainsi que la valeur des écart types (qui reflète l'étendue de la valeur des scores donnés par le correcteur) *(Tableau 1)*. Un des correcteurs a tendance à sous-coter (8 scores sur 20 sont inférieurs au score moyen de la copie – l'écart type), un autre a tendance à sur-coter (8 scores sur 20 sont supérieurs au score moyen de la copie + l'écart type). Sur les 100 scores attribués par les cinq correcteurs aux 20 copies, avec pour chacune des copies un score moyen et un écart type, on s'aperçoit qu'à 32 reprises le score attribué se situe en deçà ou au-delà du score de la copie ± l'écart type.
Le test de Kendall montre que les correcteurs sont concordants dans leur manière de classer les copies ($w = 0.597$; $X^2=56,72$; $p < 0.001$).

Résultats concernant les corrections utilisant la deuxième grille

Les cinq scores obtenus par chaque copie diffèrent en moyenne de 32 % (extrêmes 10 à 61 %). De la même façon qu'avec la première grille, ces scores varient autant pour les bonnes que pour les mauvaises copies. Le chapitre qui génère la moins grande variation est le chapitre « Matériel et Méthodes ». Le chapitre qui génère la plus grande variation de cotation est le chapitre « Résultats ».
Le *tableau 2* montre que les moyennes des scores donnés par chaque correcteur diffèrent. Les écarts types correspondants sont également très variables. Autrement dit, certains correcteurs cotent selon une échelle plus large que d'autres.
Aucun correcteur ne sous-cote ni ne sur-cote systématiquement. Cette fois-ci, aucun score n'est extérieur à l'intervalle [moyenne ± écart type] calculé pour chaque copie.

Tableau 1 :
Moyenne des scores attribués par chaque correcteur avec la première grille (Cotation sur 14)

Type	Moyenne	Ecart-type
Correcteur 1	9,2	2,76
Correcteur 2	8,65	2,35
Correcteur 3	8,15	2,06
Correcteur 4	7,4	1,90
Correcteur 5	8,85	2,18

Tableau 2 :
Moyenne des scores attribués par chaque correcteur avec la deuxième grille (Cotation sur 16)

Type	Moyenne	Ecart-type
Correcteur 1	10,6	3,03
Correcteur 2	11,85	1,76
Correcteur 3	10,6	1,88
Correcteur 4	10,95	2,09
Correcteur 5	11,35	2,08

Le test de Kendall montre que les correcteurs sont concordants dans leur manière de classer les copies (w = 0,684 ; X^2 = 64,98 ; p < 0,001).

Comparaison des résultats des corrections selon les deux grilles

En répartissant en quatre classes (< 20 %, 20 à 40 %, 40 à 60 %, > 60 %) le pourcentage de variation des cinq scores obtenus par chaque copie, on constate que les pourcentages de variation sont réduits avec la grille n°2. Ceci est illustré sur la *figure 1*. Un test de X^2 comparant ces répartitions, montre qu'elles sont statistiquement différentes avec p < 0,01.

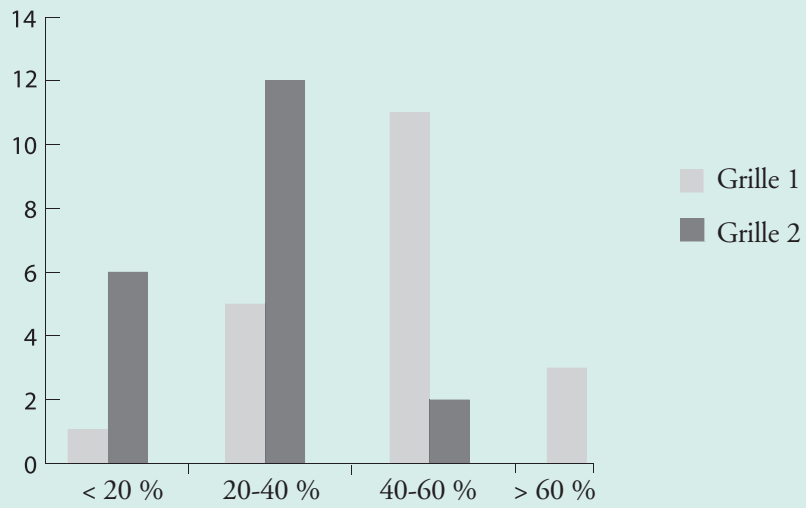
Discussion

Notre étude confirme la difficulté à mettre en place une épreuve dont la correction réponde à l'exigence métrologique de fidélité inter-correcteur pour évaluer la capacité des étudiants à analyser un article. Nous avons construit pour cette étude un protocole d'évaluation qui visait à limiter les possibilités de variation de jugement selon le correcteur : épreuve écrite de résumé, correction effectuée à l'aide d'une grille. Il apparaît que, même dans ces conditions, et même avec une grille détaillée, la cotation d'un résumé ne satisfait pas l'exigence de fidélité inter-correcteur. Cependant, nous montrons que plus la grille est détaillée, moins il y a de variation dans l'attribution du score.

Parce que la lecture critique de la littérature scientifique s'avère incontournable pour un médecin qui souhaite actualiser ses compétences, la formation à la lecture critique d'article a été intégrée au cursus médical. Les activités d'enseignement et d'apprentissage qui y concourent peuvent être mises en oeuvre dans les services lors de séances de bibliographie ou à la faculté lors de cours magistraux, de séminaires ou d'enseignements dirigés (cette dernière méthode est appliquée à Créteil). Comme tout dispositif d'évaluation, celui qui concerne l'évaluation des apprentissages relatifs à la lecture critique d'article doit répondre aux impératifs suivants : il doit être réalisable, fidèle et valide.

Le présent travail n'avait pas pour objectif d'analyser la validité de notre épreuve d'examen. En tant que caractéristique métrologique, la validité est d'ailleurs une notion complexe à appréhender. Indiquons simplement que, lorsqu'il s'agit d'évaluer des compétences et non seulement des savoirs déclaratifs, la validité dite de « construit » est tenue pour essentielle. Elle concerne la justesse et la pertinence avec lesquelles l'épreuve renseigne sur les apprentissages relatifs aux différentes composantes du « construit » théorique de la compétence visée. C'est, de loin, parmi les différents types de validité d'un examen, celui dont la démonstration est la plus complexe. Concernant la capacité des étudiants à analyser un article de manière critique,

Figure 1 :
Répartition du nombre des copies dans chacune des quatre classes de pourcentage de variation inter-correcteur de la note, selon le type de grille de correction utilisée



toute la difficulté du choix de la méthode d'examen tient au fait qu'il y a plusieurs degrés de savoirs impliqués dans cet apprentissage. A un premier niveau, il est nécessaire d'acquérir des notions d'épidémiologie, de statistiques, et de méthodologie de la recherche. Au-delà, il faut être en mesure d'utiliser ces connaissances de base pour arriver au stade de lecture critique, c'est-à-dire comprendre puis juger l'article grâce aux outils d'analyse appris. Enfin, ultime degré, il faut réussir à interpréter la littérature dans le contexte individuel du soin apporté à un patient donné, c'est-à-dire mobiliser à bon escient le savoir acquis au cours de ces lectures et enfin transmettre correctement au patient les informations nécessaires.

La lecture critique étant ainsi considérée selon plusieurs niveaux, il faut, pour déterminer le mode d'évaluation, identifier celui qui fera l'objet de l'évaluation. A cet égard, l'examen que nous avons mis en place à Créteil s'adresse à des étudiants en début d'apprentissage et évalue essentiellement l'acquisition des outils de lecture et la compréhension du texte.

La littérature concernant l'évaluation de l'apprentissage de la lecture critique d'articles est singulièrement pauvre. Il faut noter, de plus, que les articles n'explorent pas tous le même niveau taxonomique d'apprentissage. Quatre études ont utilisé des modes d'évaluation intéressants[6-9].

La première, publiée par Stern et *al.* rapporte l'utilisation d'un questionnaire à propos de l'article[6]. Pour chacun des items proposés (par exemple : « tous les devenirs importants des patients sont rapportés ») les étudiants devaient répondre selon une échelle de type Likert allant de « tout à fait d'accord » (réponse 1) à « pas d'accord du tout » (réponse 6). La même épreuve avait été préalablement proposée à cinq médecins ayant une expérience en matière de lecture critique d'article. Les réponses de ces cinq médecins avaient permis de déterminer pour chaque item une réponse « correcte », de manière à quantifier ensuite la déviation de chaque réponse de l'étudiant par rapport à cette référence. Ce type d'épreuve permet peu d'apprécier les connaissances bio-statistiques ou épidémiologiques de l'étudiant mais permet de vérifier qu'il est capable d'évaluer la qualité d'un article. Le mode de correction, dans l'optique d'un examen, est simple et rapide.

Le même principe a été utilisé par Green et *al.* qui, en plus de l'article, proposent à l'étudiant un cas clinique[7]. Par ailleurs, ils demandent à l'étudiant de justifier les réponses qu'il donne. Par exemple à la question : « les résultats de l'article peuvent être utilisés pour ce patient », l'étudiant répond sur une échelle allant de « tout à fait d'accord » à « pas du tout d'accord » et doit justifier sa réponse. La modification apportée à la méthode de Stern enrichit l'épreuve puisqu'elle permet de vérifier que l'étudiant est capable d'appliquer son nouveau savoir à un cas précis et puisque la partie rédactionnelle permet d'apprécier le raisonnement de l'étudiant. Ni la validité, ni la fidélité de ce mode d'évaluation n'ont cependant été analysées.

Plus complexe est l'organisation d'une situation simulée d'entretien avec un patient. Ce mode d'évaluation a été décrit par Bradley et Humphris[8]. L'OSCE *(Objective Structured Clinical Examination)* consistait à fournir dans un premier temps à l'étudiant deux résumés d'articles concernant la prise en charge d'une pathologie donnée. L'étudiant était ensuite mis en présence d'un « patient standardisé » interprété par un acteur professionnel préalablement préparé. L'étudiant devait ensuite réaliser un interrogatoire, un examen clinique puis expliquer au patient la prise en charge qu'il proposait. Il avait la possibilité de prescrire un traitement médicamenteux. L'acteur notait ensuite la capacité à communiquer de l'étudiant sur une échelle de type Likert en 10 points (de « très mauvais » à « très bon »). Enfin, l'acteur signalait quelle avait été la prise en charge recommandée par l'étudiant. Cette technique teste l'aptitude de l'étudiant à analyser les résumés et à mettre en pratique les résultats de cette analyse. Elle a aussi le grand avantage d'explorer ses capacités à communiquer avec le patient au sujet de ce savoir juste acquis. Cependant, ce mode d'évaluation suppose une organisation lourde et coûteuse. De plus, il avantage l'étudiant qui, au cours de ses stages hospitaliers, a eu l'occasion d'assister à des consultations dans la discipline concernée.

Une quatrième méthode, non publiée mais présentée sur internet par l'équipe du service de chirurgie de l'Université de Toronto, consiste en un examen écrit avec analyse de trois articles indépendants, concernant des domaines différents, et utilisant des méthodes différentes[9]. L'examen consiste en des questions à réponse courte avec utilisation d'échelles en sept points. Par exemple : Quels sont les critères de jugement ? Sont-ils adaptés ? Donnez une note de 1 à 7. La correction s'effectue à l'aide d'une grille. La validation de cette méthode a été réalisée lors d'une étude impliquant 44 internes. Les trois articles ont été corrigés par l'épidémiologiste et un article a été en plus corrigé par un médecin. Les conclusions de cette étude accréditent la fidélité de l'épreuve puisqu'il existe à la fois une bonne corrélation entre les scores attribués par l'épidémiologiste et ceux attribués par le médecin (fidélité inter-correcteur)

et qu'il existe une corrélation entre les scores obtenus à chacun des trois articles (consistance interne).

A Créteil, le manque de moyens et, en particulier, de temps, rend peu applicable l'organisation d'un oral, *a fortiori* un oral selon la description de Bradley et Humphris avec intervention d'un acteur[8]. Il nous a semblé qu'un examen écrit pouvait explorer les deux premières étapes de l'apprentissage, à savoir : acquisition des outils de lecture puis compréhension et critique de l'article. La structure de l'examen mis en place à Créteil a été fixée après les réflexions suivantes : premièrement, la réalisation d'un résumé est un moyen simple de vérifier que l'étudiant a compris quelle devait être la structure d'un article et quels étaient les éléments saillants et importants dans chacune des parties ; deuxièmement, les questions rédactionnelles permettent d'apprécier les acquis épidémiologiques et biostatistiques des étudiants et également la capacité à critiquer l'article à partir de ces outils. Ces deux modalités sont, en outre, expressément prévues par les dispositions réglementaires qui décrivent les épreuves de l'examen national classant français, telles qu'elles devront être mises en œuvre à partir de 2008[10]. Ce type d'épreuve (résumé et questions) vient d'être utilisé par Roussel *et al.* avec des résultats faisant suspecter une médiocre reproductibilité inter correcteurs[11]. Notre travail diffère du leur en ce que nous avons introduit une grille de correction en vue de réduire les variations de notation inter-correcteurs.

Après un an d'application, cet examen a fait l'objet d'une première évaluation dont les résultats ne sont pas présentés ici. Cette étude préliminaire a révélé qu'un étudiant performant dans les autres matières ne l'était pas forcément dans cette épreuve, que les scores obtenus au résumé n'était pas corrélés aux scores obtenus aux questions rédactionnelles et n'étaient pas reproductibles d'une session d'examen à l'autre. A l'opposé, les scores aux questions rédactionnelles des deux sessions sont corrélés. Ceci revient à dire que l'étudiant qui a réussi la première session réussit également la deuxième. De prime abord, la correction des questions rédactionnelles pose donc moins de problème. C'est pourquoi nous nous sommes concentrés dans cette étude sur le mode de correction du résumé. Il conviendra ultérieurement de s'intéresser également à la correction des questions rédactionnelles.

Le fait qu'un étudiant performant dans les autres matières puisse ne pas l'être lors de l'épreuve de lecture critique d'article peut être interprété de diverses manières, l'une d'entre elles étant que ces épreuves évaluent des caractéristiques cognitives différentes. Les autres matières donnent lieu dans notre faculté à des épreuves de résolution de cas cliniques. La mémorisation des cours est un élément déterminant pour ce type d'épreuve, alors que la réalisation du résumé puis la réponse aux questions rédactionnelles font appel à la réflexion et à des capacités d'analyse et de critique. En revanche, cela n'explique pas que les performances des étudiants aux deux sessions de lecture critique d'article de février et de juin ne soient pas corrélées. Cette discordance peut être interprétée de deux manières : elle peut rendre compte de l'inadéquation du résumé à évaluer les apprentissages concourant à l'aptitude de lecture critique (problème de validité) ; elle peut aussi résulter d'un défaut de fidélité de la correction. Notre étude avait pour objet d'apprécier la fidélité inter-correcteurs de notre mode de correction. Sa portée est limitée par le fait que les grilles ont été testées sur un seul article. Néanmoins, elle prouve l'inefficacité relative des deux grilles qui ont été proposées. Même si la grille la plus précise semble donner des résultats un peu meilleurs, il faut admettre que l'amélioration de la concordance entre correcteurs, observée avec la deuxième grille, peut être partiellement due à un « effet de pratique ». Ces deux grilles ont pourtant été utilisées dans des conditions favorables : d'une part, les correcteurs n'étaient pas des novices et avaient une bonne pratique des corrections, d'autre part, le nombre réduit de 20 copies avait été choisi de manière à limiter l'effet de lassitude qui se produit invariablement lorsqu'il y a une centaine de copies à corriger. Il est possible que certains correcteurs aient été moins rigoureux que d'autres dans l'observance de la grille ou encore que certains mots utilisés dans la grille aient donné lieu à interprétation. Une solution pour améliorer ces grilles et éviter tout phénomène d'interprétation serait, non plus de lister des items (par exemple « introduction complète et concise ») mais plutôt des mots clés dont la présence dans le résumé serait indispensable. Il faut cependant prendre garde à ne pas sanctionner, par l'usage abusif de ce système, des copies qui, bien que ne contenant pas les mots clés, témoignent cependant d'une bonne compréhension de l'article. Il y a là un risque de perdre en validité (« l'étudiant obtient une bonne note s'il a bien compris »), ce que l'on gagnerait en fidélité.

En conclusion, l'élaboration d'un résumé d'article, en tant qu'outil d'évaluation des apprentissages relatifs aux aptitudes et à la compétence de lecture critique d'un article scientifique, nous semble, dans les conditions de son développement dans notre faculté, caractérisée par une double limite : cet outil est, d'une part, incomplet et il pose pour l'instant, un réel problème de fidélité de correction. En effet, l'épreuve de résumé permet uniquement de vérifier

que l'étudiant a reconnu la structure de l'article et a compris la démarche des auteurs. Pour cette raison, le résumé devrait systématiquement être associé à un autre type d'épreuve, des questions rédactionnelles comme nous l'avons fait, et/ou des questions type Likert, qui vont, elles, explorer plutôt la capacité d'analyse et le sens critique. Par ailleurs, notre étude prouve l'existence de réelles difficultés de correction de l'épreuve de résumé. Néanmoins, avant de conclure que le résumé n'est pas une épreuve donnant des résultats fidèles, il serait licite de tester une grille encore plus précise que les deux nôtres, qui combinerait items généraux comme « l'étudiant a respecté la structure type du résumé » et présence de mots clés comme « étude prospective randomisée ».

Contributions

Christine Louis-Sylvestre a élaboré le protocole, effectué le recueil des données, participé à l'interprétation des résultats et à l'analyse statistique et rédigé les versions successives du manuscrit. Claire Furhman a participé à l'interprétation des résultats et à l'analyse statistique. Bruno Housset à participé à l'élaboration du protocole et à l'écriture des versions successives du manuscrit.

Références

1. *Linzer M, Brown JT, Fazier LM, DeLong ER, Siegel WC. Impact of a medical journal club on house-staff reading habits, knowledge, and critical appraisal skills. A randomizes control trial. JAMA 1988 ; 260 : 2537-2541.*

2. *Audet N, Gangnon R, Ladouceur R, Marcil M. L'enseignement de l'analyse critique des publications médicales scientifiques est-il efficace ? Revision des études et de leur méthodologie. Can Med Assoc J 1993 ; 148 : 945-952.*

3. *Norman GR, Shannon SI. Effectiveness of instruction in critical appraisal (evidence-based medicine) skills : a critical appraisal. Can Med Assoc J 1998 ; 158 : 177-181.*

4. *Taylor R, Reeves B, Ewings P, Binns S, Keast J, Mers R. A systematic review of the effectiveness of critical appraisal skills training for clinicians. Med Educ 2000 ; 34 : 120-125.*

5. *Parkes J, Hyde C, Deeks J, Milne R. Teaching critical appraisal skills in health care settings (Cochrane review). In : The Cochrane Library, Issue 1, 2003. Oxford : update software.*

6. *Stern DT, Linzer M, O'Sullivan PS, Weld L. Evaluating medical residents' literature-appraisal skills. Acad Med 1995 ; 70 : 152-154.*

7. *Green ML, Ellis PJ. Impact of an evidence based medicine curriculum based on adult learning theory. J Gen Intern Med 1997 ; 12 : 742-750.*

8. *Bradley P, Humphris G. Assessing the ability of medical students to apply evidence in practice: the potential of the OSCE. Med Educ 1999 ; 33 : 815-817.*

9. *McLeod R, Barkun J, Henteleff H et al. The Canadian association of general surgeons evidence-based reviews in surgery project. 2001 [On-line] Disponible sur : www.mshri.on.ca/epibiostats/images/CAGSEBRS.ppt*

10. *Conseil scientifique du centre national du concours d'internat. Epreuve écrite de lecture d'article. 2003 [On-line]. Disponible sur : http://www.cnci.univ-paris5.fr/medecine/CritiqueArticle.doc*

11. *Roussel F, Czernichow P, Lavoinne A, Lemeland JF, Fillastre JP. Reproductibilité de la correction d'une épreuve de lecture critique d'article : évaluation par une étude pilote chez 59 étudiants en médecine. Pédagogie Médicale 2005 ; 6 : 71-78.*

Manuscrit reçu le 30 novembre 2004 ; commentaires éditoriaux formulés aux auteurs le 25 mai 2005 ; accepté pour publication le 7 juin 2005.

Annexe 1 :
1ère grille utilisée pour la correction de l'épreuve de rédaction du résumé d'article

	0	1	2	3	4
Respect de la structure (introduction, matériel et méthode, résultats, conclusion)	non	oui			
INTRODUCTION comprend un bref rappel du problème énonce l'hypothèse se termine par l'objectif de l'étude	absente	incomplète	complète mais non concise	complète et concise	
MATÉRIEL ET MÉTHODES décrit les patients : 9541, critères d'inclusion décrit le critère de jugement principal décrit la structure de l'étude (prospective randomisée, double aveugle, placebo ….)	absente	absence du matériel ou de la méthode	présence du matériel et de la méthode, mais confus	complet et clair	
RÉSULTATS énonce les résultats du critère principal et des critères secondaires en donnant les chiffres avec le p	absent	aucun chiffre donné ou début d'interprétation	présence du résultat du CJP erreur sur les critères secondaires	présence du CJP seul	présence des résultats du CJP et des critères importants
CONCLUSION	absente ou erronée	non prouvée par l'article		conforme à l'article	
TOTAL					

Légende
p : valeur p (seuil de signification statistique)
CJP : critère de jugement principal

Annexe 2 :
2e grille utilisée pour la correction de l'épreuve de rédaction du résumé d'article

	0	1	2	3	4
Respect de la structure (introduction, matériel et méthodes, résultats, conclusion) (même implicite)	non	oui			
INTRODUCTION comprend un bref rappel du problème énonce l'hypothèse se termine par l'objectif de l'étude	absente	erreurs	incomplète	complète mais confuse	complète et concise
MATÉRIEL ET MÉTHODES décrit les patients : nombre, critères d'inclusion décrit le critère de jugement principal décrit la structure de l'étude (prospective randomisée, double aveugle, placebo, multicentrique, plan 2x2)	absente	absence du matériel ou de la méthode	erreurs	présence du matériel et de la méthode, mais incomplet ou confus	complet et clair
RÉSULTATS énonce les résultats du critère principal en donnant les chiffres avec le RR et le p et les résultats des critères secondaires	absent	aucun chiffre donné ou début d'interprétation	présence du résultat du CJP seul sans p ou sans RR	présence du CJP seul avec p et RR critères	présence des résultats du CJP et des secondaires importants
CONCLUSION le ranipril prévient de manière significative les complications cardiovasculaires dans une population à haut risque	absente ou erronée	non prouvée par l'article	incomplète ou confuse	conforme à l'article et complète	
TOTAL					

Légende
p : valeur p (seuil de signification statistique)
RR : Risque Relatif
CJP: critère de jugement principal