Pédagogie Médicale 2013;14 (4): 303–306
DOI: 10.1051/pmed/2013055

Un nouveau mode de correction des questions à choix multiple pourrait améliorer la fidélité et le pouvoir discriminant des épreuves classantes nationales en France

A novel scoring key for multiple choice questions could improve the reliability and discriminating ability of national ranking examinations in France

Monsieur,

Les épreuves classantes nationales (ECN) proposées en France depuis 2004 n'ont pas de pouvoir discriminant suffisant. Neuf dossiers cliniques comportant chacun quatre à 10 questions, auxquels s'ajoute l'épreuve de la lecture crique d'article (LCA) depuis 2009, ne suffisent pas à départager les étudiants puisque l'on observe de nombreux lots d'étudiants *ex æquo,* comportant parfois plus de 500 étudiants affectés du même score. Ce phénomène est mathématiquement facilement expliqué par le rapport entre le nombre d'étudiants évalués (3988 en 2004 et 7502 en 2012) et le nombre de points distribués (100 points par dossier soit un total de 1000 points). La fidélité ou reproductibilité des ECN lors d'administrations successives aux étudiants est également médiocre, comme l'a montré la comparaison entre des épreuves blanches et le classement lors des ECN officielles[1]. Le projet d'informatisation des ECN, qui seront administrées à l'aide de tablettes tactiles dans chaque faculté à partir de 2016, comporte par ailleurs la réintroduction de questions à choix multiples (QCM). Cette perspective offre une opportunité privilégiée pour réfléchir à un format de QCM qui serait susceptible d'apporter une contribution à l'amélioration des qualités psychométriques des ECN, notamment de leur capacité discriminante et de leur fidélité. De façon exploratoire, nous avons testé l'impact à cet égard d'un nouveau mode de conception et de cotation de QCM.

Un test d'évaluation des connaissances par quatre QCM a été administré à deux groupes successifs de respectivement 42 et 25 étudiants de deuxième année du deuxième cycle des études médicales inscrits au module de pédiatrie médico-chirurgicale. Les QCM ont été élaborés par deux enseignants du module de pédiatrie et relus par un groupe d'experts de la discipline, avec pour chacune d'entre elles cinq propositions de réponses différentes. Les réponses aux QCM étaient cotées selon deux modalités : 1) selon la première, dite en base 2, la réponse juste était créditée d'un point et chacune des quatre autres réponses fausses de zéro point ; 2) selon la seconde, dite en base 5, l'impact théorique de la réponse fournie sur le pronostic du patient et sa pertinence étaient pris en compte : une réponse était respectivement cotée –2 si elle devait mettre en danger le patient, –1 si elle était inutile ou dangereuse, 0 si elle n'avait aucun effet sur le pronostic du patient ou si elle était tout simplement hors de propos, +1 si elle était correcte mais insuffisante, +2 si elle était parfaitement juste (cette cote de +2 correspondant ainsi à une cote 1 dans le système

Tableau I. Exemple de question à choix multiple pouvant être cotée soit en base 2 (deuxième colonne) soit en base 5 (troisième colonne).

Quelle est la bonne proposition ?			
A	1	+2	La forme neurologique de l'invagination intestinale aigue (IIA) est rare et se manifeste comme des crises de petit mal avec absences
B	0	0	L'IIA sur diverticule de Meckel est le plus souvent réduite par lavement hydrostatique
C	0	+1	En cas de récidive d'IIA, il faut opérer
D	0	–1	La présence de sang dans la couche est un signe de gravité de l'IIA
E	0	–2	La réduction de l'IIA se fait par lavement à l'air sous contrôle échographique

A : est la bonne réponse en base 2. La réponse A est créditée du maximum de points. Elle est parfaitement juste en toutes circonstances. C'est la seule réponse qui puisse obtenir l'unanimité et le consensus parmi un groupe d'experts.
B : est une réponse fausse mais sans conséquences sur le plan pratique. En effet, une IIA développée sur diverticule de Meckel, ne peut pas être réduite par lavement. L'irréductibilité conduira inéluctablement à une décision opératoire.
C : est une réponse exacte mais à nuancer car elle ne fait pas l'unanimité. On peut dans des circonstances cliniques favorables tenter une nouvelle fois la réduction de l'IIA par lavement.
D : est une réponse fausse. La présence de rectorragie non abondante fait partie de la triade sémiologique clinique classique de l'IIA. Considérer que la rectorragie est toujours un argument de gravité conduirait l'étudiant à penser qu'il faut toujours recourir à la chirurgie de réduction de l'IIA en cas de rectorragie. Une telle décision opératoire n'est pas pertinente et ne s'inscrit pas dans la réflexion bénéfice/risque qui doit préluder à toute décision chirurgicale.
E : est une réponse fausse et correspond à une erreur grossière et une méconnaissance de base de la part de l'étudiant. L'échographie et l'air en abondance sont incompatibles.

de notation classique). Les possibilités de réponses avaient été délibérément choisies pour rendre pertinente l'application alternative des deux méthodes différentes de cotation, la catégorisation des choix ayant fait l'objet d'un consensus au sein du groupe d'experts à l'issue de nombreuses discussions ; deux réponses ne pouvaient pas mériter la même cotation. La première modalité de correction suppose que les réponses cotées zéro soient fausses sans ambigüité ; la seconde modalité introduit une hiérarchisation, aussi bien pour les deux bonnes réponses que pour les trois réponses erronées. Au total, pour l'ensemble des quatre QCM, la correction en base 2 permet de générer théoriquement cinq niveaux différents de score (de 0 à 4), tandis que la correction en base 5 permet de générer théoriquement 17 niveaux différents de score (de –8 à +8).

Un exemple de QCM élaboré selon ce format est présenté dans le tableau I. L'épreuve était anonyme ; l'enjeu de cet examen, le contexte, ainsi que les modalités alternatives de cotation avaient été expliqués aux étudiants avant l'épreuve.

L'analyse statistique, faite sur logiciel StatEL (www.adsciences.fr) a déterminé les indices descriptifs habituels (moyenne, écart-type…). Elle a recherché une corrélation entre les scores issus des deux méthodes de cotation en calculant le coefficient de corrélation par rang de Spearman ; elle a comparé les performances des étudiants selon un critère qualitatif (réussite ou non aux QCM) à l'aide du test du Chi2. La fidélité des séries de QCM a été estimée par le coefficient alpha de Cronbach, calculé sur Logiciel Excel (Microsoft) à partir des variances totales et par QCM. Le calcul du facteur d'allongement a été fait par la formule de Spearman-Brown. Le seuil de signification alpha a été fixé à 0,05.

La correction en base 2 a réparti les étudiants selon quatre niveaux de scores (de 0 à 3) au lieu des cinq théoriquement possibles, confirmant ainsi le caractère peu discriminant de la méthode de cotation. La correction en base 5 les a répartis selon neuf niveaux de scores (de –2 à +8) au lieu des 17 théoriquement possibles. L'effet « peloton », respectant une courbe gaussienne, ne disparait ainsi pas totalement

mais, avec le même nombre de questions, le nombre de niveaux de catégorisation des étudiants selon leur score est multiplié par deux. Les résultats obtenus ne sont pas différents dans les deux groupes étudiés.

Lorsque la cotation se fait en mode binaire, les étudiants se répartissent selon une courbe gaussienne centrée sur le score médian « 2 » ; la performance du groupe peut ainsi être considérée comme « moyenne ». Lorsque la cotation se fait en base 5, les étudiants se répartissent selon une courbe gaussienne centrée sur le score médian « 5 », très significativement au dessus du score moyen théorique, ce qui autorise à qualifier la performance du groupe de bonne et non plus seulement de moyenne.

La performance des étudiants, appréciée par leur score, reste stable pour sept pour cent des étudiants, quel que soit le mode de cotation ; elle augment pour 40 % et baisse pour 52 % d'entre eux. En revanche, la proportion de réussite à l'examen n'est pas différente de façon statistiquement significative (test de Chi2) selon le mode de cotation et la corrélation entre scores obtenus respectivement selon les deux modes de calcul est élevée ($r = 0,73$).

Le rang de classement est modifié pour 90 % des étudiants et les étudiants admis ne sont pas les mêmes, respectivement, dans un cas ou dans l'autre. Si la réussite des étudiants est prononcée à condition qu'ils obtiennent un score supérieur au score moyen, 64 % sont reçus lors d'une cotation en base 2 et 73 % lors d'une cotation en base 5. Si la réussite des étudiants est prononcée à condition que leur score soit égal ou supérieur à 75 % du score maximal, le pourcentage d'étudiants reçus est respectivement de 23 % et 40 %, la cotation en base 5 apparaissant ainsi plus favorable aux étudiants. Cependant, aucune de ces différences n'est statistiquement significative. Moins de 10 % des étudiants sont à la marge du seuil de réussite.

La fidélité de l'épreuve, appréciée dans la dimension de consistance interne, est faible, que ce soit lors de la correction en base 2 (coefficient alpha de Cronbach : 0,37) ou en base 5 (coefficient alpha de Cronbach : 0,25).

Au total, le nouveau mode de cotation des réponses formulées par les étudiants à une épreuve par QCM permet une meilleure distribution des performances des étudiants, sur une plage plus large de niveaux de score, sans augmenter le nombre de questions. Il ne fait pas disparaître totalement l'effet « peloton » mais semble apprécier plus positivement les performances du groupe d'étudiants. Il modifie le rang de classement des étudiants, ce qui a une incidence en cas d'évaluation à interprétation normative, ce qui est le cas dans le cadre des ECN ; en revanche, il ne modifie pas significativement le taux de réussite en cas d'évaluation à interprétation critériée.

Nos résultats ne permettent pas de documenter les impacts des réponses faites au hasard sur les scores par les étudiants, respectivement selon le mode de correction, même si on peut faire l'hypothèse que la correction en base 5 pénalise ce type de réponse, étant pris en compte, par ailleurs, qu'il n'est pas établi qu'il faille se préoccuper de façon excessive de cette problématique ; en effet, « le souci louable de chercher à corriger l'effet du choix au hasard n'a débouché jusqu'à présent que sur des résultats décevants [et] les diverses méthodes employées n'ont qu'un effet marginal sur la fidélité et la validité des mesures »[2].

Comme cela était attendu en raison du faible nombre de QCM, la fidélité de l'épreuve, appréciée dans la dimension de consistance interne, est faible, que ce soit lors de la correction en base 2 ou en base 5 ; dans le cadre limité de cette étude, il n'est ainsi pas possible de juger de la supériorité d'une méthode de cotation selon ce critère. Il est cependant intéressant de spéculer qu'avec un nombre de QCM multiplié par 4 (calculé selon la formule de Spearman-Brown) la fidélité atteindrait une valeur plus raisonnable de 0,6. En tout état de cause, on sait que le format d'administration d'un test de connaissance n'a, en lui-même, que peu d'impact sur la consistance interne du test, dès lors que la couverture des apprentissages évalués est pertinente, variée et quantitativement significative, et qu'elle atteint en pratique un niveau jugé comme acceptable (coefficient alpha de Cronbach égal ou supérieur à 0,8) à partir d'un temps d'administration de deux heures pour des QCM[3].

Il faut enfin, en conclusion, rappeler le caractère exploratoire et très préliminaire de ce travail, qui visait à examiner la possibilité d'améliorer de façon simple le caractère discriminant et la fidélité des

ECN, en se saisissant de l'opportunité de la réintroduction planifiée d'une épreuve de QCM à partir de 2016. Des études complémentaires conduites avec un plus grand nombre d'étudiants ainsi qu'avec un plus grand nombre de QCM devraient permettre de valider et renforcer nos résultats. Les difficultés de concevoir un QCM permettant une cotation selon une échelle à plusieurs points sont apparues surmontables, sous réserve d'une relecture attentive par un groupe d'experts, mais elles ne doivent pas pour autant être sous-estimées. Enfin, l'impact pédagogique des évolutions proposées, sur la nature et la qualité des apprentissages des étudiants, devra également être examiné, ce qui impliquera notamment de se préoccuper de la validité de construit des épreuves. Dans ce cadre, la place des QCM « à points variables » que nous proposons par rapport à celle d'autres tests qui n'évaluent pas les réponses des étudiants de façon binaire mais qui tentent de prendre en compte la notion d'incertitude inhérente au raisonnement clinique, tel que le permet le test de concordance de script, déjà expérimenté en France au cours du deuxième cycle[4, 5] et lui aussi introduit dans le cadre des ECN à partir de 2016, méritera d'être soigneusement étudiée et discutée.

François BECMEUR[1]
Isabelle LACREUSE[1]
Anne SCHNEIDER[1]
Philippe CLAVERT[2]
Philippe GICQUEL[1]

[1] Service de chirurgie pédiatrique, Hôpital de Hautepierre, Hôpitaux Universitaires de Strasbourg, 67098 Strasbourg Cedex, France
Mailto : francois.becmeur@gmail.com
[2] Laboratoire de pédagogie médicale, Faculté de médecine de Strasbourg, 4 rue Kirschleger, 67000 Strasbourg, France

Références

1. Vieux R, Bejot Y, Braun M, Kohler F. Capacités discriminantes et caractère prédictif d'une épreuve de type « épreuves classantes nationales » en France. Pédagogie Médicale 2011;12:159-168.
2. Gagnon R, Charlin B. Qui gagne ? Faut-il tenir compte des réponses faites au hasard au cours des examens ? Pédagogie Médicale 2007;8:69-70.
3. Van der Vleuten CPM, Schuwirth LWT. Assessing Professional competence: from methods to programmes. Med Educ 2005;39:309-17.
4. Jouneau S, Luraine R, Desrues B. Intérêt des tests de concordance de script pour évaluer le raisonnement et l'organisation des connaissances des étudiants de quatrième année des études médicales en France. Pédagogie Médicale 2012;13:225-232.
5. Caire F, Marin B, Cuny E. Utilisation du test de concordance de script au cours du deuxième cycle des études médicales : expérience dans l'enseignement de la neurochirurgie. Pédagogie Médicale 2011;2:29-35.