Qui gagne ? Faut-il tenir compte des réponses faites au hasard au cours des examens ?

Who win ? Should guessing be taken into account in examination answers ?

La question de la prise en compte de la possibilité que des réponses soient faites au hasard lors de l'évaluation des apprentissages a fait couler beaucoup d'encre au cours des trente dernières années[1].

Cette éventualité pose effectivement un problème dans un contexte académique – en Amérique du nord tout au moins – où les questions à choix multiple constituent la modalité d'évaluation dominante. Intuitivement, lorsqu'on questionne un étudiant en lui offrant un choix parmi cinq réponses possibles, on ne peut s'empêcher de penser qu'une fois toutes les cinq questions, il pourrait bien « tomber », au hasard, sur la bonne réponse, alors qu'il ne possède aucune connaissance sur le sujet. Certes, celui qui ne connaît absolument rien ne pourrait espérer au mieux qu'un score de 20 %, bien en dessous de la note de passage. Mais qu'en est-il de celui qui répondrait correctement et de manière avisée à plus de 60 % des questions mais qui répondrait avec plus ou moins de certitude, voire de manière franchement aléatoire, aux questions restantes. Quelle serait, au bout du compte, la part de devinette (« guessing ») *dans les résultats à cet ensemble de questions ?*

Plusieurs formules de correction de l'effet du « guessing » *ont été proposées au cours des années ; quelques-unes sont simples, d'autres passablement complexes. Plusieurs recherches ont été élaborées pour évaluer le gain de fidélité et de validité apporté par ces différentes approches*[2]. *Une autre approche, originale, a également été testée. Elle consiste à demander aux répondants d'estimer leur degré de confiance* (« confidence weighting » ou « confidence marking ») *pour chaque réponse donnée*[3]. *Une pondération plus ou moins complexe, en fonction du degré d'assurance, permet de corriger le score de l'étudiant, en visant à réduire de manière indirecte la part liée au hasard dans les réponses qu'il fournit. Malgré l'attrait certain de l'approche, la lourdeur de la technique qu'impose la prise en compte du degré de confiance ne semble pas s'être accompagnée de gains marqués en terme de qualités métrologiques des examens ainsi traités*[4].

Ainsi, le bilan de toutes ces tentatives amène à constater que, malheureusement, le bénéfice de solutions pourtant spontanément séduisantes n'est pas toujours confirmé lorsqu'on les confronte à l'épreuve de la réalité. Le souci louable de chercher à « corriger l'effet du choix au hasard » n'a débouché jusqu'à présent que sur des résultats décevants : les diverses méthodes employées n'ont qu'un effet marginal sur la fidélité et la validité des mesures[5]. *Ceci a conduit plusieurs auteurs à suggérer tout simplement de n'apporter aucune correction. À cet égard, la conclusion d'une revue systématique de la littérature publiée par Chevalier en 1998 est sans équivoque :* « It appears that correcting for guessing is unnecessary and should be avoided; and future studies are needed to investigate an optimal test method and a scoring formula for cognitive tests. Therefore, the conventional testing and scoring formula is recommended[6]. »

Cette conclusion, peut-être surprenante de prime abord, soulève une question importante inhérente au système des questions à choix multiple. Sans entrer dans tous les détails techniques de la logique de correction des effets du choix au hasard, on peut remarquer qu'il est d'abord assumé que la probabilité d'obtenir une bonne réponse de manière fortuite quand il y a cinq choix de réponse est d'une chance sur cinq. Mais ceci postule l'absence totale de connaissance chez l'étudiant et donc la responsabilité exclusive du hasard dans le choix des réponses. Supposons maintenant qu'un répondant possède une connaissance partielle, lui permettant d'éliminer d'emblée deux ou trois choix de réponse peu vraisemblables ; il se retrouve alors devant deux ou trois choix possibles. Imaginons maintenant que cette connaissance partielle ait comme conséquence que, pour 20 questions du test, la probabilité de choisir la bonne réponse au hasard varie respectivement entre une chance sur cinq et quatre chances sur cinq. Dans un tel contexte, les formules classiques de correction sont inadéquates, puisque que ce profil est imprévisible et

unique à chaque répondant. On peut penser que c'est pour cette raison que les formules de correction sont si peu « performantes ». Ce que l'on gagne en qualités métrologiques pour certains répondants (ceux qui répondent strictement au hasard), on le perd pour d'autres (ceux qui ont des connaissances partielles)[7].

Cependant, les chercheurs dans ce domaine ne baissent pas les bras. L'étude de Sophie Vanbelle, qui est présentée dans le présent numéro[8]*, montre que, sans chercher à corriger les effets du hasard, on peut développer une méthode originale de concordance qui permette de « dépister » des répondants qui auraient un profil de réponse particulier, peu compatible avec une réelle compétence à répondre aux questions du test. Ces répondants atypiques pourraient être évalués de façon plus approfondie pour une validation des résultats du test. Vanbelle et ses collaborateurs ont développé leur approche à partir d'un test de concordance de script (TCS) et leur étude contribue de manière significative à ce que nous connaissons de cette méthode d'évaluation, dont le système de notation tient compte de l'éventualité d'une connaissance partielle. Par ailleurs, les nouvelles approches utilisant une modélisation complexe, tels que le modèle de Rasch*[9] *– ou modèle de la « théorie de réponse aux items »* (« item response theory ») *– sont prometteuses car elles incluent dans la modélisation un paramètre de réponse au hasard. Reste à évaluer, au cours des années à venir, si ces nouveaux modèles, nettement plus complexes et donc difficiles d'approche pour les néophytes, vont apporter une valeur ajoutée significative en terme de fidélité et surtout de validité, gains qui n'ont pas été démontrés de façon claire par les méthodes utilisées jusqu'à ce jour.*

Enfin, pourquoi faudrait-il que les statistiques aient toujours le dernier mot ? Plusieurs auteurs suggèrent que la meilleure façon de contourner le problème serait d'élaborer des tests mieux construits. Pas si bête comme idée!

Robert GAGNON, Bernard CHARLIN
Centre de pédagogie appliquée aux sciences de la santé (CPASS)
Faculté de médecine, Université de Montréal (Québec, Canada)
Mailto:rgagnon@cmq.org ;
bernard.charlin@umontreal.ca

Références

1. *Nunnally JC. Psychometric Theory. New York : McGraw Hill, 1978.*
2. *Choppin BH. Correction for guessing. In: Keeves JP (Ed.). Educational research, methodology, and measurement: an international handbook. Oxford: Pergamon Press, 1988:384-6.*
3. *Leclercq D. Confidence marking, its use in testing. In:Choppin B & Postlethwaite N (Eds.). Evaluation in Education: International Review Series, Oxford : Pergamon, 1982:6:161-287.*
4. *Crocker L, Algina J. Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart & Winston, 1986.*
5. *Thomas J, Prihoda TJ, Pinckard RN, McMahan CA, Jones AC. Correcting for Guessing Increases Validity in Multiple-Choice Examinations in an Oral and Maxillofacial Pathology Course. J Dent Educ 2006;70:378-86.*
6. *Chevalier SA. A Review of Scoring Algorithms for Ability and Aptitude Tests. Paper presented at the Annual Meeting of the Southwestern Psychological Association, New Orleans (LA), April 1998.*
7. *Downing S. Guessing on selected-response examinations. Med Educ 2003;37;670.*
8. *Vanbelle S, Massart V, Giet D, Albert A. Test de concordance de script : un nouveau mode d'établissement des scores limitant l'effet du hasard. Pédagogie Médicale 2007;8:71-81.*
9. *Rasch G: Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960 (Réédité en 1980).*