

Comment améliorer la qualité des examens QCM à l'université sur base d'une analyse des items ? L'exemple de l'UCL

Résumé court

Le service d'évaluation en appui de la qualité (EVA) offre aux enseignants de l'UCL une aide logistique à l'évaluation des acquis des étudiants par QCM. Le système informatisé de conception et de correction des examens qui a été choisi, ConTest (Consilium Benelux, 2012), permet de réaliser différentes analyses statistiques des résultats : calcul d'un coefficient de fidélité (alpha de Cronbach), étude des indices de difficulté et de discrimination, relation entre la difficulté et le pouvoir discriminant de l'item, caractéristiques de la distribution de fréquence des leurres, etc. Pour chaque examen, les résultats de ces analyses d'items sont présentés dans un rapport qui est transmis aux enseignants. Le service EVA leur offre un accompagnement individuel qui les aide à interpréter et à exploiter ces rapports afin d'améliorer la qualité de leur QCM et des items qui le composent. Au niveau de l'ensemble de l'université, nous avons rassemblé actuellement des données relatives à plus de 500 examens. On observe une grande variabilité des pratiques en ce qui concerne le nombre d'items proposés par les enseignants, le nombre d'options de réponse par question, les barèmes de notation, l'utilisation ou non d'une pénalisation en cas de mauvaise réponse, etc. Le croisement de ces variables avec les indicateurs psychométriques issus des analyses d'items (alpha de Cronbach, indice de discrimination, indice de difficulté) permet de formuler des balises et des recommandations qui devraient aider les enseignants à améliorer la fidélité et la validité de leurs QCM.

Mots clé : QCM, analyse d'items, docimologie, validité, fidélité

Résumé long.

Ces dernières années, la massification des étudiants dans certaines filières d'études a fait exploser le nombre d'examens et de copies à corriger. Un nombre de plus en plus important d'enseignants se sont dès lors tournés vers le Questionnaire à Choix Multiple (QCM) pour évaluer et certifier les acquis de leurs étudiants.

Les QCM ont cependant une mauvaise réputation en milieu universitaire. Leur commodité de correction les rend attractifs mais selon certains auteurs ils n'inciteraient qu'à poser des questions sur des détails de seconde zone (Leclercq 2006). Une autre faiblesse attribuée aux QCM est de laisser la place au hasard en permettant à l'étudiant de sélectionner la bonne réponse sans la connaître.

Mais les QCM ne présentent pas que des inconvénients ! Un de ses principaux intérêts est de pouvoir couvrir une grande étendue de la matière. Ce type d'examen donne en effet la possibilité de poser un nombre élevé de questions, ce qui n'est pas le cas des autres formes d'évaluation (examen oral et examen écrit à questions ouvertes). Si elles sont bien conçues, ces questions constitueront un bon échantillonnage de la matière, l'examen sera ainsi plus représentatif et il permettra d'assurer une objectivité et une fiabilité

¹ Service d'évaluation en appui de la qualité (EVA) de l'Université catholique de Louvain (UCL)

suffisantes à l'évaluation (Considine & Botti, 2005, Downing, 2004, Laveault & Grégoire, 1997).

Cette question est essentielle car “sans fiabilité, toute discussion sur la validité devient futile” (Laveault & Grégoire, p. 137). A cet égard, le développement récent de systèmes de traitements automatisés par lecture optique présente un intérêt majeur car les analyses d'items qu'ils réalisent permettent d'évaluer la fiabilité de l'examen et d'identifier les items qui posent problème en vue d'aider l'enseignant à prendre les décisions nécessaires pour les améliorer.

Le service d'évaluation en appui de la qualité de l'UCL a acquis récemment un de ces outils. Ce système appelé ConTest (Consilium Benelux, 2012) permet de créer des grilles d'examens, d'en réaliser la lecture optique, d'effectuer leur correction automatisée et de produire un rapport avec les résultats des étudiants. Il permet aussi de réaliser une analyse statistique des données. Parmi les différents indicateurs proposés par le système, les 3 suivants sont analysés en priorité par notre service.

- **L'alpha de Cronbach.** Ce coefficient traduit la “cohérence interne” de l'épreuve et sert d'indice pour estimer la fiabilité de l'examen. Plus ce coefficient est élevé (et se rapproche de 1,00), plus nous pouvons postuler que les items mesurent effectivement la même caractéristique ou le même objectif et contribuent par conséquent à produire des résultats fiables. Si l'alpha de Cronbach est faible, le score obtenu par l'étudiant ne permet pas de prédire avec suffisamment de précision quel est son niveau réel de connaissances.
- **L'indice de facilité** d'un item se définit comme la proportion (p) d'étudiant ayant coché la réponse correcte parmi les différentes options proposées. Plus cet indice est élevé (ex. 0.89) plus l'item est facile. Plus cet indice est faible (ex. 0,17), plus la question est difficile. Le système ConTest calcule un indice corrigé (p') qui tient compte de l'effet du hasard. Dans un QCM, il faut en effet tenir compte de la probabilité de réussir les questions sans vraiment connaître la réponse. ConTest calcule un indice de facilité pour chaque item et pour l'ensemble de l'examen (moyenne des p' des items).
- **L'indice de discrimination** (r_{ir}) est un indice important car il permet de vérifier que la question joue bien son rôle dans le cadre d'un examen : départager correctement les étudiants qui ont eu un score total élevé à l'examen des étudiants qui ont eu un score total faible. L'indice de discrimination repose sur le calcul d'une corrélation entre la performance de l'étudiant à chaque item et sa performance aux autres questions de l'épreuve. Une corrélation dont la valeur se situe à 0,20 et plus indique que les étudiants ayant réussi cet item s'avèrent avoir aussi obtenu des notes élevées aux autres items. Un indice de discrimination se rapprochant de 0,00 indique que la question n'a aucun pouvoir discriminatif. Une corrélation inférieure à 0,00 est problématique : les étudiants qui ont une note basse au test réussissent l'item alors que ceux qui ont une note élevée l'échouent. Il faut donc éliminer de tels items. ConTest calcule un indice de discrimination pour chaque item et pour l'ensemble de l'examen (moyenne des r_{ir} des items).

Un graphe permet de visualiser la “place” qu'occupe chaque item de l'examen en fonction de son pouvoir discriminant (valeur r_{ir} en abscisse) et de sa difficulté (valeur p' en

ordonnée). Il est complété par une analyse de la distribution de fréquence des leurres. On s'attend à ce qu'un bon distracteur soit sélectionné par des étudiants peu performants et qu'il soit ignoré par les "bons". Des distracteurs qui ne sont jamais choisis devraient être éliminés. Si certains leurres sont choisis plus souvent que la bonne réponse par des bons étudiants, cela pose question.

Chaque enseignant s'adressant au service EVA reçoit ces rapports d'analyses. Un accompagnement individuel à la demande est organisé pour les aider à les interpréter et à les exploiter. Mon QCM a-t-il été bien construit ? Dois-je augmenter le nombre de questions ? Puis-je, au contraire, me permettre de les réduire ? L'examen est-il trop simple ou trop difficile ? Les scores obtenus sont-ils fiables ? Chaque leurre a-t-il été efficace ? Quels items pourrais-je conserver pour un autre examen sur le même contenu ? Quels sont les items à éliminer ?

Une banque de données compile l'ensemble des paramètres relatifs aux 500 QCM qui sont passés ces deux dernières années par le système ConTest. On observe une grande variabilité des pratiques en ce qui concerne le nombre d'items proposés par les enseignants, le nombre d'options de réponse par question, les barèmes de notation, l'utilisation ou non d'une pénalisation en cas de mauvaise réponse, etc. Le croisement de ces variables avec les valeurs statistiques de l'analyse d'items (alpha, p', rir) permet de dresser une liste de recommandations qui devraient aider à l'avenir les enseignants à améliorer la fidélité et la validité des leurs QCM.

Bibliographie

- Considine, J. & Botti, M. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education, *Collegian*, 12, 19-24.
- Consilium Benelux (2012). ConTest Statistics. Version 0.19. Mars 2012.
- Downing, S.M (2004). Reliability : on the reproductibility of assesment data. *Medical Education*, 38, 1006-1012.
- Ebel, R.L. (1969). Expected reliability as a function of choices per item, *Educational and Psychological Measurement*, 29, 565-570.
- Girard, M. & Normand, S. *Guide de lecture d'un rapport d'analyse d'items*. Université de Montréal, Bureau d'évaluation, Faculté de médecine, 2001 (<http://www.medbev.umontreal.ca/docimo/DocSource/Guide%20de%20lecture%20AnItem.pdf>)
-
- Grier, J.B. (1975). The number of alternatives for optimum test reliability. *Journal of educational measurement*, 12, 2, 109-113.
- Lord, M. (1977). Optimal number of choices per item. A comparison of four approaches. *Journal of educational measurement*. 14, 1. pp. 33-38.
- Laveault, D. & Grégoire, J (1997). *Introduction aux théories des tests en sciences humaines*. Paris, Bruxelles : De Boeck Université & Larcier.
- Leclercq, D. (2006). L'évolution des QCM. In G. Figari et L. Mottiez-lopez. *Recherches sur l'évaluation en éducation* (pp. 139-146). Paris : L'Harmattan.
- Morissette, D. (1997). *Guide pratique de l'évaluation sommative : gestion des épreuves et des examens*. Paris, Bruxelles : De Boeck & Larcier .
- Normand, S. (2001). *L'analyse d'items, une technique docimologique pour valider des questions d'examen*, Diaporama du mode d'emploi du logiciel AnItem, Université de Montréal, Bureau d'évaluation, Faculté de médecine. Montréal. (www.medbev.umontreal.ca/docimo/DocSource/Docimologie_AnItem.pps)
- Revue Réseau (2009). *Le questionnaire à choix multiple*. Facultés Universitaires Notre-Dame de la Paix. Namur (<http://www.unamur.be/pdf/publications/67449.pdf>).

- Ricci J.-L. (2010) *ABC des questionnaires à choix multiples*. Ecole polytechnique fédérale de Lausanne, Unité de Recherche et appui pour la formation et ses technologies (CRAFT). Lausanne. (<http://teaching.epfl.ch/webdav/site/craft/shared/import/ABC%20des%20QCM.pdf>).