Recyclage de questions à choix multiples d'épreuves certificatives de physique en items de tests formatifs en ligne

Marique, P.-X.[1], Van de Poël, J.-F.[2], Hoebeke, M.[3]

[1] pxmarique@ulg.ac.be ; Didactique de la Physique, Département de Physique, Université de Liège

[2] jfvandepoel@ulg.ac.be ; IFRES-eCampus, Université de Liège

[3] m.hoebeke@ulg.ac.be ; Didactique de la Physique, Département de Physique, Université de Liège

Depuis septembre 2012, en Fédération Wallonie-Bruxelles de Belgique, les étudiants inscrits en première année de bachelier en médecine sont interrogés sous forme de questions à choix multiples lors des épreuves certificatives des cours scientifiques donnés au premier quadrimestre de leurs études. A l'Université de Liège, les examens du cours de physique sont organisés en deux parties composées chacune d'environ 25 questions. Les étudiants ont la possibilité de présenter cette épreuve à maximum 3 reprises. Chaque année académique, approximativement 150 questions inédites sont donc créées.

Ces questions sont répertoriées, et éventuellement améliorées, afin d'être réutilisées dans le cadre d'un outil d'entraînement en ligne, appelé « simulateur d'examens » (Marique, 2014, 2015). Cet outil est destiné à améliorer des connaissances directement en lien avec une stimulation constante de la mémoire des étudiants à l'aide de QCM (Roediger, 2006, 2011).

Même si lors de leur création, les 20 règles de conception développées par Leclercq (Leclercq, 1986) sont bien suivies, l'analyse a posteriori des questions d'examens permet leur amélioration en vue de leur introduction dans l'outil en ligne. A cet effet, chaque item est testé à l'aide de plusieurs indices.

- L'indice de difficulté P informe sur la proportion d'étudiants ayant correctement répondu à la question. Plus l'indice P est faible, plus la question sera jugée difficile. A l'inverse, plus l'indice P est élevé, plus la question est supposée facile. La valeur « corrigée » P', tenant compte d'un éventuel effet de hasard, est également calculée ((Laveault & Grégoire, 1997).
- Le coefficient de discrimination D d'un item, servant à en déterminer la validité, est également mesuré. En effet, « un "bon" item est un item qui serait réussi par une plus grande proportion de sujets ayant obtenu un score élevé à l'examen que par des sujets ayant obtenu un score faible » (Laveault & Grégoire, 1997, p. 231). Autrement dit, ce coefficient devrait permettre de répondre à la double question suivante : les étudiants qui réussissent cet item sont également ceux qui ont obtenu un score élevé sur l'ensemble de l'épreuve ? (Kelley, 1939 ; Findley, 1956, Ebel ; 1965, Gilles, 2002) Les étudiants qui ont échoué à cet item font-ils également partie du groupe des scores les plus faibles sur l'ensemble de l'épreuve ? Pour ce faire, deux groupes sont isolés de la population totale : un premier groupe composé des 27 % des étudiants ayant obtenu les meilleures notes au test et un second reprenant les 27 % des étudiants caractérisés par les plus faibles notes. Le nombre d'étudiants ayant réussi l'item dans chaque groupe est alors mesuré. La différence divisée par le nombre d'individus de chaque groupe fournit la valeur du coefficient de discrimination.

En cas de valeurs non satisfaisantes pour ces deux indices, la question problématique est alors étudiée afin de déterminer ce qui a pu poser problème et éventuellement non retenue pour l'entraînement des étudiants. Après modification, elle est intégrée à l'outil en ligne. Les questions utilisées en ligne seront également ré-évaluées à intervalle régulier (1 fois par an).

L'ensemble des questions composant le simulateur d'examens est ensuite évalué par plusieurs membres de l'équipe pédagogique du Département de Physique de l'Université de Liège. Chacun attribue un niveau de difficulté à chaque question en fonction de 4 critères empiriques : abstraction, réflexion, mixité, mathématiques. Chaque évaluateur attribue une note de 1 à 3 à chacun des critères. La somme obtenue indiquera le niveau de difficulté total de la question selon la règle suivante : [4 ; 5] => 1 ; [6 ; 7] => 2 ; [8 ; 9] => 3 ; [10 ; 12] => 4.

Le critère d'abstraction est lié à la simplicité/complexité de l'énoncé, au fait que les données sont implicites/explicites, … L'indice de réflexion est conditionné entre autres par le nombre d'étapes à effectuer pour parvenir à la solution, à la quantité de passage d'une formule à l'autre dans le raisonnement. La mixité sera simplement évaluée en fonction du nombre de matières, chapitres, compétences à mobiliser pour résoudre la question. Enfin, la valeur octroyée au critère mathématique dépendra de la quantité de formules et relations à utiliser, des compétences à mobiliser en algèbre et en analyse, … L'attribution des valeurs pour chacun de ces critères est purement subjective. Cependant, afin d'assurer et de vérifier la concordance entre les évaluateurs, le kappa de Fleiss (Fleiss, 1981) a été mesuré pour chaque épreuve et nous indique, en général, une valeur approximative de 0,4, jugée satisfaisante.

Les questions sont alors classées par pool dépendant de leur niveau de difficulté et de la matière abordée. Elles permettront alors à l'étudiant un travail progressif sur le simulateur d'examens.

A l'avenir, dans une perspective d'amélioration du classement par niveau des questions disponibles dans le simulateur d'examens, la concordance entre l'indice P' d'un item, dépendant des résultats des étudiants à celui-ci, et le niveau de difficulté qui lui a été attribué par l'équipe pédagogique, sera étudiée.

L'impact de ce système sur l'apprentissage des étudiants fait actuellement l'objet d'un suivi particulier et précis. Toutes les tentatives des étudiants et leurs résultats sont enregistrés systématiquement en prévision d'une analyse des effets de ce système.

Références :

BRENNAN, R.-L., *A generalized upper-lower item discrimination index,* Educational and psychological measurement, 32, 289-303, 1972

EBEL, R.L.,*Confidence–Weighting and Test Reliability*, Journal of Educational Measurement, 2,49-57
B., 195

FINDLEY, W.G., *A rationale for evaluation of item discrimination statistics*, Educational and Psychological Measurement, 16, 175-180, 1956

FLEISS, J.-L., *Statistical methods for rates and proportions*, 2nd ed., John Wiley, New York, 1981

GILLES, J.-L., *Qualité spectrale des tests standardisés universitaires,* Thèse de doctorat, Université de Liège, 2002

KELLEY, T.L., *Selection of upper and lower groups for the validation of test items*, Journal of

Educational Psychology, 30, 17-24, 1939

LAVEAULT, D.,GREGOIRE, J. *Introduction aux théories des tests en sciences humaines*. Paris, Bruxelles : De Boeck Université & Larcier s.a., 1997

LECLERCQ, D. *La conception des questions à choix multiple*, Bruxelles, Labor, 1986

LECLERCQ, D. *Qualité des questions et signification des scores avec application aux QCM*, Bruxelles, Labor, 1987

MARIQUE, P.-X., HOEBEKE, M. *Plate-forme interactive au service des grandes populations d'étudiants suivant un cours de Physique,* Actes de la Conférence TICE 2014, Béziers, France, 2014

MARIQUE, P.-X., VAN DE POËL, J.-F., HOEBEKE, M. *Quel outil d'entraînement pour des étudiants en médecine évalués par QCM en physique ?,* Actes du Colloque ADMEE 2015, Lisbonne, Portugal, 2015

ROEDIGER, H. L., BUTLER, A. C. *The critical role of retrieval practice in long–term retention*, *Trends in Cognitive Sciences*, volume 15, n°1, pp. 20–27, 2011. http://dx.doi.org/10.1016/j.tics.2010.09.003

ROEDIGER, H. L., KARPIPCKE, J. D., *The power of testing memory: Basic research and implications for educational practice*, Perspectives on Psychological Science, volume 1, n°3, pp. 181–210, 2006. http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x

VIAU, R. *La motivation en contexte scolaire,* St-Laurent, Éditions du Renouveau pédagogique, 1994

https://www.uclouvain.be/cps/ucl/doc/adef/documents/EVA_QCM_version3.pdf, consulté le 24/09/2016