hep/

Haute école pédagogique
Recherche et développement
Relations nationales et internationales
Avenue de Cour 33 — CH 1014 Lausanne
www.hepl.ch

Conditions méthodologiques d'utilisation des degrés de certitude pour l'auto-estimation des compétences dans le cadre du test de maîtrise de français EFES à la HEP Vaud

Note préparée par Jean-Luc Gilles à l'attention de la CONFAC du 22 novembre 2012

Depuis plusieurs décennies la plupart des institutions d'enseignement supérieur européennes sont confrontées à une forte augmentation de leurs effectifs d'étudiants alors que les budgets alloués n'augmentent pas en proportion (Gibbs & Jenkins, 1992). Cette situation entraîne dans les sections où les étudiants sont les plus nombreux, notamment les premiers cycles d'études, un recours de plus en plus massif aux épreuves standardisées avec questionnaires à choix multiple, ce qui permet d'évaluer dans des délais raisonnables de grands groupes d'étudiants (De Landsheere, 1979 ; Leclercq 1986 ; Gilles, 2002 ; Blais et Gilles, 2011).

La technique des degrés de certitude fait partie des outils à la disposition des évaluateurs qui ont recours aux épreuves standardisées. De nombreuses recherches en édumétrie ont démontré son utilité, citons notamment, et à titre d'exemple : De Finetti (1965) ; Shufford & al. (1966) ; Leclercq (1982, 1993, 2003) ; Leclercq et Bruno (1993) ; Bruno, J. (1993) ; Leclercq et Gilles (2001) ; Gilles (1997, 2010).

En invitant l'étudiant à accompagner son choix d'une proposition à une QCM du pourcentage de chances qu'il lui attribue d'être correcte (son degré de certitude) nous permettons plus de nuances dans l'analyse de ses performances. A un extrême, le choix d'un distracteur accompagné du pourcentage de certitude maximum (100%) présente la pire des situations, celle où l'étudiant fournit une réponse erronée en estimant qu'elle a un maximum de chances d'être correcte. A l'opposé, l'étudiant qui répond correctement avec une certitude maximale fait preuve d'une connaissance assurée. Entre ces deux extrêmes, s'ouvre tout l'espace d'une analyse « spectrale » (et non plus « binaire ») des performances, espace « invisible » lorsque les pourcentages de certitude ne sont pas utilisés. Ainsi, dans le cas d'une réponse correcte, Jans & Leclercq (1999) proposent une terminologie *ad hoc* pour distinguer une « ignorance » (réponse correcte et certitude faible), d'une « connaissance partielle » (réponse correcte et certitude moyenne), d'une « connaissance assurée » (réponse correcte et certitude élevée). De telles nuances spectrales ont aussi été envisagées par ces auteurs dans le cas d'une réponse incorrecte (« méprise » et « connaissance dangereuse »).

La technique des degrés de certitude associée aux questions à choix multiple permet donc de dépasser le caractère « binaire » de l'évaluation des performances des étudiants (la proposition choisie est soit correcte, soit incorrecte), mais à condition de veiller à respecter une série de règles méthodologiques que Shufford & al. (1966) appellent « admissible probability measurement procedures ».

D'autres enjeux liés à l'utilisation des degrés de certitude sont présentés sur une page en annexe de ce document.

Vous trouverez dans les pages qui suivent une courte note de synthèse qui résume les conditions méthodologiques qui doivent être rencontrées pour garantir un recueil sans biais des données liées à l'auto-estimation de ses compétences à l'aide de la technique des degrés de certitude. Le lecteur trouvera des informations plus complètes à ce propos dans Leclercq (1993, pp. 141-143 - http://hdl.handle.net/2268/18634) et Leclercq et Gilles (2001, pp. 134-146 - http://hdl.handle.net/2268/4828).

(1) Le barème de tarifs doit être conforme à la théorie des décisions

Il s'agit de gratifier une réponse correcte accompagnée d'un degré de certitude élevé d'un meilleur score que si elle était accompagnée d'une certitude faible et inversement pour les réponses incorrectes. Les tarifs du barème des points doivent être calculés de manière à favoriser une seule stratégie : celle qui consiste à dire la vérité. Le barème des points ci-après garantit que l'expression de son intime conviction rapporte plus de points que tout autre stratégie.

(2) La consigne doit être "probabiliste".

Demander à l'étudiant d'indiquer sa certitude par des termes vagues du type "peu sûr", "moyennement sûr", "très sûr", etc. est à proscrire car ces expressions recouvrent des réalités différentes en fonction des sujets. De plus, avec des termes aussi flous la variabilité est telle chez un même étudiant qu'on ne peut même pas recourir à des traitements ordinaux intra-sujets. Voici la consigne mise au point par Leclercq (1983, 1993, 1995) et qui est actuellement utilisée dans le cadre du test de maîtrise du français proposé par le groupe Evaluation du Français pour l'Enseignement Supérieur (EFES) (Defays, 2010).

Si vous considérez que votre réponse a une probabilité d'être correcte comprise entre	Ecrivez	Vous obtiendrez les points suivants en cas de réponse correcte (RC)	Vous obtiendrez les points suivants en cas de réponse incorrecte (RI)
0 % et 25 %	0	+ 13	+ 4
25 % et 50 %	1	+ 16	+ 3
50 % et 70 %	2	+ 17	+ 2
70 % et 85 %	3	+ 18	+ 0
85 % et 95 %	4	+ 19	- 6
95 % et 100 %	5	+ 20	- 20

Les coupures sur l'axe ne sont pas équidistantes ce qui permet une expression du degré de certitude plus nuancée à l'extrémité supérieure de l'échelle. Ainsi, l'étudiant peut faire la distinction entre 90 % (valeur centrale de la certitude 4) et 97,5 % (valeur centrale de la certitude 5) bien que la différence soit de 7,5 % seulement. Dans le premier cas (90 %) il n'a qu'une chance sur dix (1/10) de se tromper tandis que dans le second (97,5 %) il n'a qu'une chance sur quarante (1/40), soit 4 fois moins. Etablir la même différence au milieu de l'échelle, par exemple entre 40 % (1/1,7) et 47,5 % (1/1,9), n'est pas pertinent car nous ne sommes pas capable de distinguer ces deux derniers « rapports »...

(3) Le calcul d'indices métacognitifs doit être possible

La consigne utilisée autorise le calcul d'un indice de réalisme basé sur les différences entre les taux d'exactitude et les valeurs centrales des intervalles de probabilité ainsi que le calcul d'un indice de centration basé sur la différence entre la certitude moyenne et le taux d'exactitude moyen et dont le signe détermine la surestimation (+) ou la sous-estimation (-). Il existe différentes variantes pour la formule de réalisme (Lichtenstein & al., 1975 ; Leclercq, 1975, 1983 ; Leclercq & al., 1993). Dans le cadre du test de maîtrise du français nous avons utilisé la formule de Leclercq adaptée par Gilles (2002, 2010) afin que le minimum de réalisme soit égal à 0 et le maximum égal à 1.

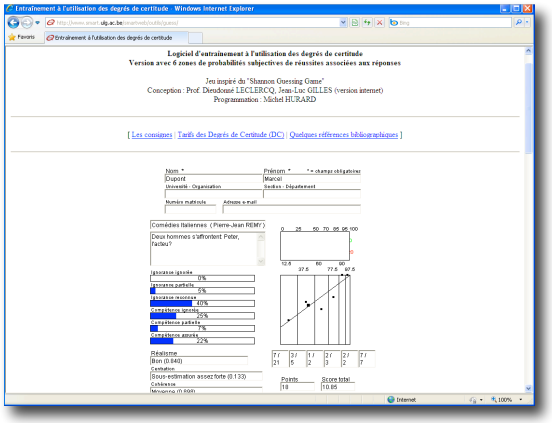
(4) Les sujets doivent être préalablement entraînés à l'utilisation des degrés de certitude

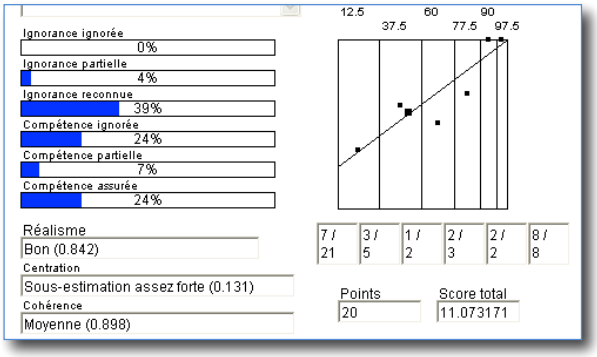
Il est conseillé d'informer et d'entraîner les étudiants avant d'employer les degrés de certitude tels que décrits ci-avant en vue de les familiariser avec leur utilisation. Différentes formules existent : l'accès via internet d'informations concernant les consignes, des simulations d'épreuves corrigées automatiquement en ligne, des logiciels d'entraînement tels que GUESS, etc.

L'entraînement GUESS[1] a été décrit en détail par Leclercq & Gilles (1994). La version en ligne propose à chaque étudiant un entraînement à l'utilisation des degrés de certitude à l'aide d'un jeu où il s'agit de

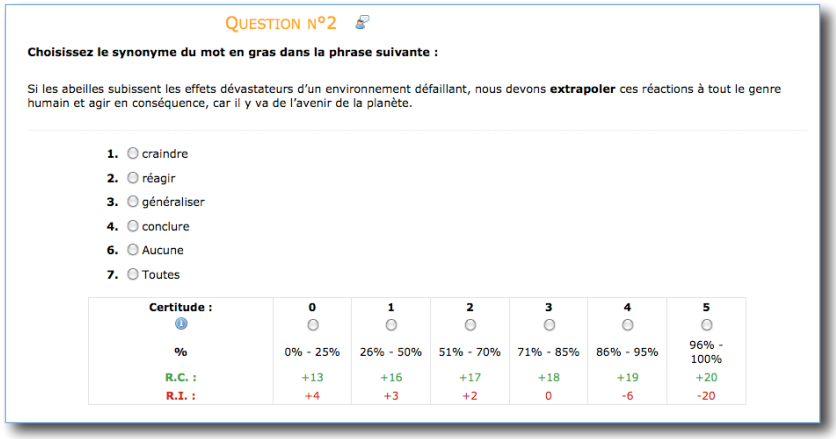
[1] La version en ligne du logiciel GUESS est accessible à l'adresse : www.smart.ulg.ac.be/smartweb/outils/guess

deviner les lettres successives d'un texte d'au moins cent lettres (jeu inspiré des travaux de Shannon, 1951 & d'Attneave, 1959). Le joueur effectue une prédiction en tapant une lettre qu'il accompagne de la probabilité subjective de réussite exprimée à l'aide d'un degré de certitude. Il est ensuite informé de la réponse correcte qui s'affiche dans la zone réservée au texte. Lettre par lettre, le texte s'affiche ainsi à l'écran. Evidemment, le début des mots est plus difficile à deviner que leur fin (c'est ce que Shannon voulait démontrer). Après un nombre donné de réponses un graphique de réalisme se construit dans le coin supérieur droit de l'écran, l'étudiant peut y observer la justesse des probabilités subjectives de réussites qu'il a attribuées à ses prédictions (en abscisse) en les confrontant aux taux de réussite (en ordonnée) pour chaque degré de certitude.





La simulation en ligne de l'épreuve permet d'entraîner les étudiants en leur proposant des questions du même type que celles qui figureront dans le test.



Ces deux types d'entraînement, avec chacun leurs avantages et leurs inconvénients, se complètent bien. GUESS permet de donner beaucoup de réponses accompagnées de certitudes en peu de temps et de visualiser un feedback immédiat personnalisé. Cependant, l'exercice proposé est assez éloigné de la situation du questionnaire EFES à choix multiple en quatre parties : vocabulaire, orthographe, syntaxe et compréhension. A l'inverse, la simulation en ligne ne permet pas de poser beaucoup de questions, par contre celles-ci y sont similaires à celles qui sont posées à l'examen, tant du point de vue de la forme que du contenu.

Bibliographie

Attneave, F. (1959). Application of information theory to psychology. New York: Holt, Rinehart and Winston.

Blais, J.-G. (2008). Evaluation des apprentissages et technologies de l'information et de la communication – Enjeux, applications et modèles de mesure – Tome I. Québec : Presses de l'Université de Laval.

Blais, J.-G. et Gilles, J.-L. (2011). Evaluation des apprentissages et technologies de l'information et de la communication – Le futur est à notre porte – Tome II. Québec : Presses de l'Université de Laval.

Bruno, J. (1993). Using testing to provide feedback to support instruction: a reexamination of the role of assessment in educational organizations. NATO ASI Series, Item Banking: Interactive Testing and Self Assessment, Berlin: Springer Verlag, 1993, Vol. 112, pp. 190-209.

De Landsheere, G., (1979). Dictionnaire de l'évaluation et de la recherche en éducation, Presses universitaire de France, Paris.

Defays, J.-M. (2010). Actes de la journée d'étude du groupe EFES, Université de Liège, 27 février 2010.

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item, British Journal of Mathematical and Statistical Psychology 18, pp. 87-123.

Gilles, J.-L. (1997). Impact de deux entraînements à l'utilisation des degrés de certitude chez les étudiants de 1ère candidature de la Faculté de Psychologie et des Sciences de l'Education de l'ULg. In Boxus, E. (ed.), Actes du 15e colloque AIPU, Liège, 6-10 juillet 1997. Liège : Affaires Académiques de l'Université de Liège, pp. 311-326.

Gilles, J.-L. (2002). Qualité spectrale des tests standardisés universitaires – Mise au point d'indices édumétriques d'analyse de la qualité spectrale des évaluations des acquis des étudiants universitaires et application aux épreuves MOHICAN check up '99, thèse de doctorat en Sciences de l'éducation. Liège : Université de Liège, Faculté de psychologie et des sciences de l'éducation de l'université de Liège (559 pages).

Gilles, J.-L. (2010). Qualité spectrale des tests standardisés universitaires. Sarrebruck : Editions universitaires européennes

Gilles, J.-L., Detroz, P., Crahay, V., Tinnirello, S. et Bonnet, P. (2011). La plateforme ExAMS, un "assessment management system" pour instrumenter la construction et la gestion qualité des évaluations des apprentissages. Dans Blais, J.-G. et Gilles, J.-L. (Dir.), Evaluation des apprentissages et technologies de l'information et de la communication – Tome II. Québec : Presses de l'Université de Laval.

Hunt, D. (1993). Theory and application to learning and testing. NATO ASI Series, Item Banking: Interactive Testing and Self Assessment, Berlin: Springer Verlag, 1993, Vol. 112, pp. 177-189.

Leclercq, D. (1975). L'évaluation subjective de la probabilité d'exactitude des réponses en situation pédagogique.Thèse de doctorat en Sciences de l'Education, Université de Liège Institut de Psychologie et des Sciences de l'Education.

Leclercq, D. (1982). Confidence marking, its use in testing. Postlethwaite, Choppin (eds.) Evaluation in Education, Oxford : Pergamon, 1982, vol. 6, 2, pp. 161-287.

Leclercq, D. et Gilles, J.-L. (2001). Techniques de mesure dans l'autoévaluation. Dans G. Figari & M. Achouche (dir.), L'activité évaluative réinterrogée – Regards scolaires et socioprofessionnels. Bruxelles : Editions De Boeck, pp. 134-142.

Leclercq, D. et Gilles J.-L. (1994). GUESS, un logiciel pour entraîner à l'auto-estimation de sa compétence cognitive. Actes du colloque QCM et questionnaires fermés, Paris: ESIEE, 1994.

Leclercq, D. (1986). La conception des questions à choix multiple, Bruxelles, Ed. Labor.

Leclercq, D. et Bruno, J. (1993) Item banking: interactive testing and self assessment : proceedings of the NATO advanced research workshop held in Liege, Belgium, October 27-31, 1992. Berlin: Springer Verlag.

Leclercq, D. (1993). Validity, Reliability, and Acuity of Self-Assessment in Educational Testing. NATO ASI Series, Item Banking: Interactive Testing and Self Assessment, Berlin: Springer Verlag, 1993, Vol. 112, pp. 114-130.

Leclercq, D. & al (1993). The Taste approach: General implicit solutions in MCQq, open books exams and interactive testing and sel-assessment. NATO ASI Series, Item Banking: Interactive Testing and Self Assessment, Berlin: Springer Verlag, 1993, Vol. 112, pp. 210-232.

Leclercq, D. (2003). Diagnostic cognitif et métacognitif au seuil de l'Université : le projet Mohican mené par les 9 universités de la Communauté française Wallonie-Bruxelles. Editions de l'Université de Liège.

Lichtenstein et al. (1975). Calibration of probabilities : the state of the art, decision making and change in human affairs Proceedings of the Fifth Research Conference on Subjective Probability, Utility and Decision Making, Darmstadt, 1-4 September, D. Reidel.

Shufford, E. et al (1966). Admissible probability measurement procedures. Psychometrika 31, pp.125-145.

Annexe

Les enjeux liés à l'utilisation des degrés de certitude

La connaissance n'est pas affaire de tout ou rien

Dans le contexte des évaluations ayant recours aux questionnaires à choix multiple, on considère habituellement une réponse fournie à une question de façon binaire : soit elle est correcte, soit elle est incorrecte, sans se préoccuper de nuances liées à la conviction avec laquelle l'étudiant a répondu. Les différents états de connaissance partielle qui découlent de l'association d'une réponse et d'un degré de certitude autorisent un diagnostic plus subtil et par là différents niveaux de remédiation. Cette amélioration de la sensibilité de l'outil d'évaluation contribue également à une mesure plus subtile des modifications intra-individuelles.

Le doute est le moteur même de la connaissance

La prise de conscience de son incompétence, de son incertitude favorise chez l'apprenant une rupture d'équilibre qui peut l'amener à rechercher l'information, à interroger son environnement afin de réduire cette incertitude.

La production de jugements est un des niveaux d'objectif les plus élevés...

...et, paradoxalement, des moins évalués ! Par exemple, la taxonomie d'objectifs pédagogiques de BLOOM (1956) propose au sommet de la hiérarchie le niveau « évaluation » qui comprend la production de jugements qualitatifs ou quantitatifs. Force est de constater que ce niveau taxonomique n'est guère entraîné et évalué.

L'incompétence est une situation normale de la vie

Les domaines dans lesquels chacun de nous est compétent sont bien moins nombreux que ceux où il est ignorant et ... *il faut vivre avec...* Amener les étudiants à prendre conscience de leur degré d'incompétence peut les aider à augmenter leur niveau de compétence.

L'ignorance reconnue n'est pas dangereuse

L'ignorance avouée n'a pas de conséquence sociale négative, par contre, l'ignorance ignorée, elle, est dangereuse ! Mieux vaut ne pas s'improviser médecin, pharmacien, secouriste, pilote d'avion, ... et reconnaître les limites de ses compétences.

L'ignorance dissimulée est dangereuse

Habituellement, on considère qu'il est honteux de ne pas savoir. Cependant, dans maintes situations, c'est le fait que des personnes aient tenté de dissimuler leur ignorance qui provoque des catastrophes, et non le fait d'avoir avoué son incompétence.

L'auto-évaluation s'apprend par l'expérience personnelle

Il n'y a pas, à notre connaissance, de règles et principes d'auto-estimation de ses compétences qu'on puisse enseigner. Par contre, l'apprentissage de cette habileté métacognitive se fait par l'ajustement de nos comportements d'auto-estimation après avoir été confronté aux conséquences de nos jugements (d'où l'importance d'associer aux degrés de certitude un barème de tarifs conforme à la théorie des décisions).

Pour en finir avec la correction for guessing...

Citons sur ce point De Landsheere (1979, p. 76) : "*La correction pour divination fait l'objet de nombreuses critiques. Elle repose notamment sur l'hypothèse gratuite que tous les sujets ont également deviné. De plus, on n'établit pas la distinction entre l'élimination de certains choix sur base de connaissances réelles et la divination au pur hasard. Une correction beaucoup plus adéquate est assurée quand le sujet indique dans quelle mesure il est certain de sa réponse*". Leclercq (1988, p. 306) cite cinq raisons d'abandonner la *correction for guessing* classique et de la remplacer par les degrés de certitude: (1) elle basée sur un modèle théorique faux, (2) elle est injuste, (3) elle n'est pas formative, (4) elle n'est pas informative, et enfin, (5) elle est restreinte aux QCM.