Systèmes d'évaluation innovants et critères de qualité

Dieudonné Leclercq
Université de liège – LabSET – d.leclercq@ulg.ac.be

Jeudi 18 Mai 2006 de 10h20 à 10h50 Amphithéatre des thèses

Les dernières décennies ont été riches en matière d'innovations. Les lister sans plus serait déjà édifiant. Nous souhaitons cependant ici en situer certaines dans le cadre de critères de qualité d'une évaluation (Leclercq, 2005).

A. D'un test à un Système

Quand Carver (1974) a défini les deux types de propriétés des **tests**, à savoir leurs propriétés psychométriques d'une part et édumétriques de l'autre, il est resté assez vague. Bien sûr, il a attribué aux **propriétés psychométriques** la capacité de fournir des mesures valides, fidèles et sensibles des différences entre les individus, bref de rendre compte des variations interindividuelles. Symétriquement, il a attribué aux **propriétés édumétriques** la capacité de rendre compte de manière valide, fidèle et sensible des modifications intraindividuelles, des changements, cognitifs ou affectifs ou sensori-moteurs ou encore relationnels par exemple, chez une même personne, soit dans le temps pour un même type de performance et de contenu (on parle alors de gains et de perte, ou de modification qualitative), soit pour des performances et des contenus différents (on parle alors de flexibilité de stratégies d'apprentissage, de profil adaptatif personnel), et, dans les deux cas, d'ambivalence et de polyvalence mathétiques[1].

Nous avons (Leclercq, 2005, chap 1, « La rose des vents des fonctions et caractéristiques des évaluations pédagogiques ») décrit les **propriétés** des évaluations, les **caractéristiques** des mesures et l'**utilisation** de l'information qui en est faite, mais sans porter de jugement de qualité. Un tel jugement n'a, en effet, de sens que quand ces caractéristiques sont combinées en un système, donc en rapport à une axiologie, une épistémologie, une stratégie.

Le titre de Carver se concentrait sur le concept de test. Nous partons d'un concept apparemment plus complexe, parce qu'il se veut plus analytique : celui de Système d'Evaluation de Performances d'Individus en Apprentissage. Par « d'individus », nous voulons exclure de notre propos les évaluations des institutions ou des enseignements, même si celles-ci ont bien des points communs avec l'évaluation des étudiants. Nous parlons de « performances », permettant de faire, éventuellement des inférences sur les compétences (qui assurent la reproductibilité des performances). Enfin, et surtout, nous parlons de **système,** dont la définition classique est « un ensemble d'éléments (ici de caractéristiques et de qualités) interdépendants ». Ainsi, les scores des étudiants dépendent évidemment de leurs réponses qui, elles-mêmes, dépendent des consignes (y compris les modes et barèmes de correction) et des types de questions (QROC, QCM, QRM, SGI, DC[2]) qui, elles-mêmes dépendent des objectifs ou visées (formative ? certificative ?) de l'évaluation, qui elle-même dépend de l'axiologie de la formation (à court terme ? à long terme ?)..
Nous pensons que les systèmes d'évaluation des performances d'individus en apprentissage (SEPIA) peuvent être **décrits** par « la rose des vents » de l'évaluation et **évalués** (en qualités) par la grille ETIC PRAD que nous développerons ci-après. Par le pluriel au mot « qualités », nous voulons indiquer qu'il va falloir prendre en compte plusieurs critères, et que la qualité ne se résumera pas à UN nombre ou UNE position sur une échelle.

Notre grille ETIC PRAD s'est inspirée de l'approche VENTURE (Validity, Examinee appropriateness, Normed excellence, Teaching feedback, Usability, Ethics) développée par le

[1] Du verbe « manthanau » (en grec ancien) : « j'apprends ».

[2] Nous aurons l'occasion ci-après de définir chacune de ces expressions.

Centre for the Study of Evaluation de UCLA et appliquée à des milliers de tests. (Leclercq, 2005, 5, 35). Ce qui importe ici est d'illustrer la multiplicité des critères et la complexité de certains d'entre eux (la validité par exemple).

B.. Les 8 qualités ETIC PRAD d'un SEPIA

B1. Validité Ecologique

Cette expression est due à Egon Brunswick (1943). Elle signifie la mesure dans laquelle l'évaluation respecte les conditions naturelles d'exercice de la performance. La plupart des tests en amphithéâtre ont une faible validité écologique car, pour des médecins, par exemple, les « cas réels » n'apparaissent pas sur papier, mais en chair et en os. De même, pour mesurer la capacité de manœuvrer chez un conducteur de camion, lui demander d'introduire au clavier une grandeur angulaire de braquage ne rencontre pas le critère de validité écologique car, dans la réalité, c'est au moyen d'un volant et à 1,5m du sol qu'il aura à manœuvrer. Certaines personnes ont en effet de grandes capacités de « manœuvres dans l'espace », mais dans certaines situations.

Ceci rejoint l'idée de Howard Gardner (1996), le promoteur de l'idée des « intelligences multiples » qui note *« La mesure d'une intelligence donnée ... devrait mettre en lumière les problèmes susceptibles d'être résolus dans les données et les outils propres à cette intelligence »*....(1996, 48) et *« Quand les individus sont évalués dans des conditions proches de « véritables situations de travail », il est possible de prédire leur résultat final avec beaucoup plus de précision »* (1996,158)

Quand nous défendons (Leclercq, 1983, 1993, 2003) le recours aux Degrés de Certitude (DC) acompagnant une réponse, c'est entre autres parce que nous pensons que ce procédé a une plus grande validité écologique que le testing habituel qui empêche les étudiants d'exprimer leur doute. Choppin (1975) a décrit ce problème dans ses modèles 1, 2 et 3. Il dénonce la vision manichéenne (tout ou rien) de phrases telles que « répondez uniquement si vous savez ; omettez si vous ne savez pas », alors que nous sommes très souvent (et en particulier lors de situations d'apprentissage) dans des états de connaissance partielle. (DeFinetti, 1965). Des sentiments du genre «si je pouvais, j'irai relire le cours » ou « j'irais voir dans le dictionnaire », sont résumés dans le Degré de Certitude, comme nous l'avons montré expérimentalement (Leclercq & Boskin, 1990).

Plus les arguments seront puisés dans la vie courante, plus l'épreuve aura une « validité apparente » (en anglais *face validity*).

B2. Validité Théorique.

Elle prend deux grandes formes : validité de contenu et validité de construct. .

La première est la validité de **contenu** (ou de « couverture » du contenu) : tout ce qu'il faut tester l'est-il et rien que cela ? On résout ce problème en « équilibrant les questions » issues de divers chapitres ou de divers domaines. Les Progress Tests de Maastricht (250 questions tous les 3 mois sur toute la médecine) garantissent que chacun des grands domaines (la criculation, la respiration, la reproduction, etc.) sont chacun représentés par un nombre suffisant de questions dans le test (Leclercq et Vandevleuten, 1998, 200).
La seconde forme de validité théorique est souvent appelée validité de **construct** (introduite par Cronbach et Meehl en 1955): Le système d'évaluation des performances cognitives (SEPIA) est-il fondé sur un modèle crédible (scientifiquement fondé) des **Processus Mentaux** ? Le test correspond-il à la théorie concernant la variable mesurée ? Les auteurs d'un test doivent établir ce type de validité par des arguments empruntés aux grandes théories et par des résultats expérimentaux jugés fiables.
Les 6 niveaux de la taxonomie des objectifs cognitifs de Bloom, qui correspondent plutôt à des « processus mentaux » sont une base robuste (utilisée et adaptée dans de très nombreux contextes), mais qui demande des affinages théoriques et techniques. Ainsi, il nous paraît important de distinguer

la connaissance « de recognition » de celle e' « évocation ». Les travaux de Luh (1922), Bahrick (1984), Shruwirtz (2000) confirment que les taux de réussite dans des épreuves de remémoration sont radicalement différentes selon la consigne et la technique utilisée.

B3. Validité Informative (ou Diagnostique)

Partons d'un exemple célèbre : l'examen oral. Potentiellement, il sollicite chez l'étudiant à peu près tous les niveaux de la taxonomie des processus cognitifs de Bloom évoquée en B2 ci-avant. Cependant, il est fréquent que la communication vers l'étudiant se résume à un « c'est satisfaisant », car plusieurs indices concordants (qui se consolident l'un l'autre) confortent l'interrogateur dans son jugement. L'interrogateur peut aboutir à cette intime conviction sur base de « configuration de réponses » très différentes : la faiblesse dans une facette peut être compensée par une force dans une autre.

Au terme d'un examen oral, cependant, il est quasiment impossible, et cela a été montré expérimentalement (Englehart, 1994 ; Leclercq, 2005, 3, 26), de donner un score pour chacun des 6 niveaux de la taxonomie de Bloom. Et ce, entre autres, parce que l'interrogateur n'a pu observer qu'un échantillon de certains, parfois même aucun échantillon, mais un « composé » combinant diverses dimensions. Sur chacune d'entre elles, il ne peut pas, sur la base des données à sa disposition, donner un diagnostic sûr, ni souvent un diagnostic du tout.

Les QCM classiques (une seule solution est correcte et il faut la choisir), dont on connaît les faiblesses (Leclercq, 1986, 30-42) peuvent être largement améliorées par le recours à des variantes, par exemple les Solutions Générales Implicites (solutions Aucune, Toutes, Manque de données et Absurdité), appliquées en « cascade » et permettre ainsi de distinguer entre compréhension, application et analyse (Leclercq, 1993 ; Leclercq, 2005,4,27-29).

Les Degrés de Certitude (appliqués à des QCM ou à des Questions Ouvertes) permettent, quant à eux, de mesurer, séparément et en plus des 5 premiers niveaux de la taxonomie de Bloom, le 6°, à savoir l'évaluation – jugement, et ce par des indices ad hoc tels que la Confiance et l'Imprudence. (Leclercq et Poumay, 2005). Baragabiribije (2003) a montré (en physique en première année universitaire) que les Degrés de Certitude étaient un révélateur des « idées fausses » ou « misconceptions » dans une population d'étudiants. Dans les réponses à de telles questions, l'indice d'Imprudence (certitude moyenne accompagnant les réponses incorrectes) est supérieur à l'indice de Confiance (certitude moyenne accompagnant les réponses correctes), ce qui est anormal

B4. Validité Conséquentielle

Attribué à Samuel Messick (1984), ce concept porte sur les conséquences sociales d'une évaluation. La validité conséquentielle d'une évaluation s'apprécie aux suites que cette évaluation a sur les représentations, les actes (ex : réviser ou nonb la matière, changer ou non de méthode d'étude) des apprenants, des formateurs ou d'autres personnes.

Les test classiques d'intelligence, dans la mesure où ils ne parvenaient pas à être « culture free » oint contribué à accréditer l'idée de « races supérieures ». Des informations édumétriques peuvent être incompréhensibles, ou non utilisables et donc sans conséquences pour les apprenants. C'est ce que nous avons nous-même vécu avec les indices de réalisme (par calibration par exemple) inexploitables dans le concret pour permettre à l'apprenant de se déterminer pour une stratégie d'étude. Nous avons dès lors opté pour d'autres indices, nouveaux : Confiance et Imprudence. L'expérience RESSAC (Résultats d'Epreuves Standardisées au Service de l'Apprentissage en Candi) a montré que la moitié des étudiants qui ont reçu a temps une information diagnostique sur leurs résultats (différence entre scoire en mémorisation et en compréhension) ont changé leur stratégie d'étude...et oint eu un taux de réussite nettement supérieur (Leclercq et al, 2003, 155-167)

Autre exemple : les étudiants de 1° année universitaire ont tendance à surestimer leur compréhension de mots d'un texte technique. Nous leur avons démontré sur leurs propres performances. De nombreux

étudiants ont dit avoir, en conséquence, consulté beaucoup plus souvent le dictionnaire, avec efficacité, ce qu'un post-test a confirmé (Leclercq et al., 2002).

POUR APPROFONDIR LE CONCEPT : DES ASPECTS CONSÉQUENTIELS DE LA VALIDITÉ, VOIR GREEN, 1998 ; LINN, 1998 ; LUNE, PARKE, & STONE, 1998 ; MOSS, 1998 ; RECKASE, 1998 ; TALEPOROS, 1998 ET YEN, 1998.

B5. Validité Prédictive (ou concurrente)

Les mesures obtenues permettent-elles de prédire efficacement (c'est-à-dire avec précision) d'autres mesures ? Par exemple celles obtenues par une procédure ou un test beaucoup plus coûteux en temps, en spécialistes, en argent, etc. Ou ces mesures permettent-elles de prédire des événements ultérieurs, ar exemple la réussite scolaire plusieurs mois plus tard, ou la réussite professionnelle plusieurs années plus tard, etc. ? A nouveau, c'est la corrélation entre les mesures prédictives et les mesures critères (ou à prédire) qui permet de répondre à cette préoccupation.
Le cas échéant, la validité prédictive peut être établie en l'absence de validité de « construct »; c'est le cas lorsqu'un instrument prédit efficacement sans que l'on comprenne pourquoi. Ce type de situation n'est pas propre à l'éducation.
Par exemple, dans l'étude MOHICAN (Monitoring Historique des Candidatures), on a calculé la Corrélation de chacun des 10 tests passés aux 4000 entrants dans les universités de la Communauté française de Belgique et la réussite de ces étudiants 12 mois plus tard (Leclercq, 2003, 146-149). Evidemment, les épreuves les plus prédictives pour une faculté (ex : philosophie et Lettres) ne sont pas les mêmes que pour une autre (par exemple médecine vétérinaire).

B6. Replicabilité – Fidélité (*Reliability*)

Cette qualité est la stabilité de la mesure : réitérée, donnerait-elle les mêmes valeurs ? Par exemple, une copie corrigée par un autre examinateur obtiendrait-elle la même note ? C'est une question classique en docimologie critique (Leclercq, 2005,ch. 3). Les formes parallèles d'un test standardisé donnent-elles la même note ? Le sous-test constitué des questions paires d'une épreuve et le sous-test constitué des questions impaires donnent-ils les mêmes notes ? On a souvent tendance à penser qu'un test fidèle est forcément valide. Il n'en est rien. Crocker and Algina (1986, page 217), cités par Aces (2006) démontrent la différence entre la fidélité et la validité avec l'analogie suivante :

"*Consider the analogy of a car's fuel gauge which systematically registers one-quarter higher than the actual level of fuel in the gas tank. If repeated readings are taken under the same conditions, the gauge will yield consistent (reliable) measurements, but the inference about the amount of fuel in the tank is faulty*".

Une formule (Spearman Brown) détermine le nombre de questions et le nombre de distracteurs nécessaires pour obtenir un niveau de fidélité donné (0,8 par exemple).

On peut se poser la question de la façon inverse : quel doit être le coefficient d'allongement n du test pour atteindre une fidélité donnée (par exemple 0,80 ou 0,90) d'un test qui existe déjà et dont on connaît la fidélité actuelle ? On répond aussi par une formule (Guilford et Fruchter, 1978, p. 432) à cette question.

B7. L'Acceptabilité - Applicabilité

Pour le professeur, le problème d'**adhésion** peut se poser en termes divers.
Au niveau **axiologique**, certains enseignants considèrent que leurs évaluations doivent servir plus à sélectionner qu'à former. Il est vrai que les enseignants doivent faire les deux. Mais dans quelles proportions ?

Au niveau **épistémologique**, certains enseignants considèrent que la capacité d'évaluation (le niveau le plus élevé de la taxonomie des objectifs cognitifs de Bloom) doit intervenir dans la notation des performances des étudiants, par exemple en accordant des points (supplémentaires, donc positifs !) au réalisme dans l'auto-évaluation des compétences. Ce réalisme peut être confronté à la réalité et on peut calculer objectivement la surévaluation et la sous-évaluation. Bruno De Finette (1965) avait cependant, il y a 40 ans déjà, montré que « *Seule la probabilité subjective peut donner un sens objectif à toute méthode de notation et à tout score en résultant.* ». D'autres enseignants, cependant, n'acceptent pas, par principe, de combiner deux mesures dont une fait appel à la subjectivité de l'apprenant (pourtant mesurée objectivement).

Au niveau technique, certains enseignants sont allergiques à certaines formes d'évaluation. Les uns aux QCM (dont ils ne connaissent souvent que la forme la plus débile : les QCM classiques) , les autres aux réponses ouvertes, posant des problèmes de concordance de notation interjuges (et même intrajuges).

Au niveau **pratique**, l'applicabilité est soubent facile à définir en termes de durée, de matériel nécessaire (ordinateur ou seulement papier ?), de disponibilité de nombreux spécialistes le meme jour au même moment, de précautions antifraude, de dimension ou de nombre de locaux, de moment de la journée ou de la semaine, etc. Certains de ces critères sont des sources de rejet de procédures d'évaluation.

Pour l'étudiant, les problèmes d'**adhésion** se posent aussi. Ainsi, il arrive fréquemment que des aides soit offertes aux étudiants, par exemple de passer des tests formatifs pour se faire une idée de leur niveau de compétence, mais que peu d'étudiants profitent pas de ces offres, soit parce qu'elles interfèrent avec des horaires déjà chargés, parce que ces aides sont trop peu personnalisées (en contenu, en intensité, etc.) ou pour d'autres raisons encore. C'est pourquoi certains enseignants essayent des procédures alternatives ; comme l'illustrent Hanzen et al. (2006) en recourant au testing formatif à distance.

Plus **la familiarité** de l'étudiant avec les procédures d'examen, avec les barèmes de cotation, etc. plus il est « aguerri aux tests », en anglais « *test wiseness* » (Leclercq, 1986, 108-109), plus ses chances de réussite sont élevées. C'est une des raisons pour lesquelles les étudiants sont souvent méfiants devant de nouvelles procédures d'évaluation, et ils ont raison de l'être quand celles-ci aboutissent à des scores qui leur sont plus défavorables.

B8. Déontologie - Ethique

Certains types d'épreuves créent des injustices **inter-étudiants**. Ainsi, les examens oraux favorisent les « extravertis », les écrits ceux qui ont « une belle écriture » ou « une bonne orthographe ». Certaines épreuves à temps limité (speed tests) défavorisent les « réflexifs » (selon l'expression de Kagan, 1955), etc.

Des règlements d'examen assurent la **transparence** des procédures : droit de voir les notations du professeur sur sa copie, droit à des demandes de recorrection, recalculabilité des scores, contrôlabilité, imputabilité, doubles corrections (en Tunisie par exemple).

-libre des discordances de jugement inter-juges (docimologie négative). Le sérieux dans la prévention ou la répression de la fraude fait aussi partie des qualités déontologiques d'un examen, de même qu'une certaine discrétion dans la communication des résultats désastreux.

C. Des exemples d'innovations en termes d'ETIC PRAD

Exemple 1.

Face à des grands groupes (plus de 300 étudiants), j'ai d'abord recouru aux **QCM** simples, tout en ayant des regret quant à divers aspects « qualité » de cette procédure. Au fil des ans, j'ai ajouté des caractéristiques (Leclercq, 1986, 2005).

D'abord les Solutions Générales Implicites et les Degrés de Certitude qui permettent d'améliorer la **validité théorique** (plus de niveaux de la taxonomie de Bloom pris en compte) et la **prédictivité**.
Ensuite, la constitution de deux types d'examen : à Livres Ouverts (autres niveaux de la taxonomie cognitive), et à Livres Fermés (épreuve de mémoire) via des QCL, cette dernière technique permettant d'améliorer la **validité de contenu** par le grand nombre de questions possibles corrigées par ordinateur.
Ensuite encore, les questionnaires spectraux améliorant **l'acceptabilité** (car correction immédiate et calcul des indices métacognitifs sur le champ) et le debriefing immédiat (communication des réponses correctes, discussion sur ces réponses) améliore la **validité informative** (les étudiants savent exactement et de suite quelles questions ils ont réussies ou échouées et pourquoi).
L'insistance sur des questions « pièges » augmente la **validité conséquentielle** en amenant les étudiants à lire le livre plus en profondeur, à plus poser de questions et surtout à lire beaucxoup plus attentivement les questions de l'examen.
Enfin, très récemment, augmentation de la **validité théorique** par des Questions Ouvertes écrites permettant de tester le niveau « Synthèse - Expression » de la taxonomie de Bloom que nous ne mesurions plus depuis que, il a deux ans, nous avions abandonné les examens oraux (pour des raisons d'**acceptabilité** : cela nous prenait trop de temps. Poser deux questions ouvertes au lieu d'une vise à augmenter la **replicabilité** des scores.
Toutes ces dispositions ne parviennent cependant pas à garantir une **validité écologique** de ce type d'examen, et c'est surtout regrettable en fin de cursus, dans la phase de formation la plus professionnalisante, donc, par exemple à l'agrégation (où les étudiants ayant leur Maîtrise se forme pour pouvoir enseigner dans le secondaire). Ce n'est pas étonnant que ce soit pour ces étudiants que nous avons conçu un testing informatisé par cas (Delcomminette et Leclercq, 2006).
La préparation systématique à ces procédures via des « répétitions à blanc », ou tests formatifs, nous semble contribuer à la **validité déontologique** de notre dispositif, notre SEPIA (Leclercq, 2005).

Exemple 2.
Les **Tests de Progression**, uniques au monde, appliqués depuis 30 ans en médecine à l'université de Maastricht (Vandervleuten & Wijnen, 1990 ; Leclercq et Vandervleuten, 1998 ; Verhoeven, 2003), consistent à imposer aux étudiants, quelle que soit l'année (des 6) qu'ils fréquentent, à passer un même test (250 Vrai-Faux) sur toute la matière de médecine, et ce quatre fois par an. Cela impose à une équipe spécialisée de construire des formes parallèles tous les 3 mois, représentatives de toutes les composantes de la matière (**validité de contenu**), en nombre de questions suffisant (250 !), d'une durée **acceptable** (3 à 4 heures), et à haute **validité informative** ou diagnostique (voir un exemple de feedback personnalisé dans Leclercq et Vandervleuten, 1998, 200-201). Cependant, ces *Progress tests* manquent de **validité écologique**.

Exemple 3.
C'est pourquoi à Maastricht, les Tests de Progression sont de plus en plus complémentés par des **ECOS** (Examens Cliniques Objectifs Structurés), concept créé par Harden (Harden et al., 1975 ; Harden, 1988). Dans l'application qu'en font Bourguignon et al. (1997) en pédiatrie, ce type d'examen exige 14 locaux d'examen médical dans un couloir d'hôpital. Dans chaque local, un interrogateur jouant d'abord le rôle de patient simulé, puis de notateur (sur base d'une grille d'évaluation pondérée). La durée d'interaction ente l'étudiant en médecine et chaque évaluateur dure 6 minutes au terme desquelles, au signal d'une sonnerie, l'étudiant change de local (appelé « station »). Le jeu de rôle ainsi assumé tant par l'interrogateur que par l'interrogé, la pression du temps, la variété des « cas » contribuent à la **validité écologique**, ce qui est reconnu par les étudiants.

Exemple 4
En pédagogie, nous avons développé la stratégie de formation / évaluation appelée **PARM** ou Projets d'Animations Réciproques Multimédias (Jans et al., 1998 ; Leclercq, Marotte et al. , 2003). Cette méthode consiste à « faire donner le cours par les étudiants ». Cette définition est légèrement exagérée. En tout cas, par équipes (de 2 à 5 étudiants), il y a animation (à l'aide de PowerPoint, de vidéos, d'activités.) du grand groupe. Cette formule vise à être en concordance avec des objectifs transversaux (**validité de construct**) tels que la conception d'une communication / animation, la recherche de

documents et l'approfondissement d'une matière en petit groupe, la répartition de tâches, la réalisation multimédias, etc. Une grille de notation (10 critères annoncés) est remplie séance tenante par quatre évaluateurs (**fidélité de notation**). Cette grille tient compte des objectifs annoncés (voir ci-dessus), et les notes sont communiquées et commentées par écrit au groupe (**validité informative**). Les étudiants sont invités à préciser quels ont été les investissements de chaque membre du groupe dans quelle tâche, mais indépendamment des juges qui ignorent, au moment où ils fixent leur note, lequel des étudiants est responsable de quel aspect du travail., et l'inverse est vrai pour les étudiants (**validité déontologique**).

D. Conclusion

Ce n'est pas l'innovation en évaluation en elle-même qui est intéressante, mais sa contribution à l'amélioration d'une ou plusieurs critères de qualité. Certaines innovations peuvent faire gagner en qualité sur certains aspects et perdre sur d'autres. L'approche « qualités » est donc forcément plurielle. Elle n'est interprétable que dans un Système d'Evaluation d'Individus en Apprentissage (SEPIA).

Bibliographie

Aces : Validity Handbook.
http://www.collegeboard.com/highered/apr/aces/vhandbook/evidence.html

consulté le 15-4-2006.

Bahrick, H.P. (1984). Semantic memory content in permastore : 50 years of memory for Spanish learned in school. Journal of Experimental Psychology : General, 120, 1-29.

Baragabirije, D. (2003). Contribution à la realisation et à la mise en ligne d'activités de remediation en physique générale. Mémoire de DES en Technologie de l'Education et de la Formation. Liège : ULg ; Namur : FUNDP.

Bourguignon, J-P, Albert, A. & Senterre J., Apport d'une évaluation de type E.C.O.S. (Evaluation Clinique Objective et Structurée) en 6e année de médecine, in Boxus, E., Jans, V., Gilles, J.L. & Leclercq, D. (Eds) Stratégies et médias pédagogiques pour l'apprentissage et l'évaluation dans l'enseignement supérieur, Actes du 15e colloque de l'Association Internationale de Pédagogie Universitaire (AIPU), Liège : STE-Affaires Académiques, 1997, 245-246.

Brunswick, E. (1943) Organismic achievement and Environment Probability, Psychological Review, 50, 255-272.

Carver, R.P. (1974), Two dimensions of tests : Psychometric and edumetric, American Psychologist, 29, 512-518.

Choppin, B. (1975), Guessing the Answer on Objective Tests, Brit. Journal of Educational Psychology, 45, 206-213.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. Belmont, CA : Wadsworth.

Cronbach, L. J. and Meehl, P. M. "Construct Validity in Psychological Tests," Psychological Bulletin 52 (1955): 281-302.

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

Delcomminette, S. et Leclercq, D. (2006). Transférer de la théorie à la pratique via un Testing Informatisé par Cas en Formation d'Enseignants (TICAFE). 23° Colloque de l'AIPU, Monastir.

Englehard, G. JR. (1994), Examining rater errors in the assessment of written composition with a many-faceted Rasch model, Journal of Educational Measurement, summer 1994, vol. 31, 2, 93-113.

Gardner, H. (1996). Les intelligences multiples, Paris : Retz, trad de Multiple intelligences. The theory in practice. A reader. (1993) Basic Books.

Glass, G.V. & Stanley, J.C. (1970). Statistical Methods in Education and Psychology. Englewood Cliffs, NJ: Prentice-Hall.

Glass, G.V.; McGaw, B. & Smith, M.L. (1981). Meta-analysis in Social Research. Beverly Hills, CA: SAGE, 279 pp.

Glass, G.V.; Cahen, L.S.; Smith, M.L. & Filby, N.N. (1982). School Class Size: Research and Policy. Beverly Hills, CA : SAGE

Glass (cours en ligne): http://glass.ed.asu.edu/stats/lesson0/

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. Educational Measurement, 17, 16-19, 34.

Hanzen, Ch., Pluvinage, P. et Leclercq, D. (2006). Impact de tests formatives organises via la plateforme WebCT sur la réussite aux tests certificatifs. 23° Colloque de l'AIPU, Monastir.

Harden, R.M., Stevenson, M., Downie, WW, Wilson, G.M. (1975). Assessment of clinical competence using objective structured examination. Br Med, 1, 447-451.

Harden, R.M. (1988). What is an OSCE ? Med Teach, 10, 19-22.

Jans, V., Leclercq, D., Denis, B. & Poumay, M. (1998) Projets d'Animations Réciproques Multimédias (PARMs), in D. Leclercq (Ed). Pour une pédagogie universitaire de qualité. Sprimont : Mardaga, 207-241.

Kagan, J. (1965c), Impulsive and reflective children : significance of conceptual tempo. In Krumboltz (Ed.), Learning and the Educational Process, Chicago, Rand Mc Nally, 133-161.

Leclercq D., (1982) Confidence marking. Its use in Testing, in Postlethwaite & Choppin (Eds) , Evaluation in Education, Oxford : Pergamon, vol 6, 2, 161-287.

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

Leclercq, D. (1987), Qualité des questions et signification des scores, Bruxelles : Labor.

Leclercq, D. & Boskin, A. (1990), Note taking behavior studied with the help of hypermedia, in Estes, Heene & Leclercq (Eds), Proceedings of the 7th International Conference on Technology and Education, Bruxelles, mars 1990, 2, pp. 16-19.

Leclercq D. & Bruno J. (1993), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin : Springer Verlag.

Leclercq, D. & Van der Vleuten, C. (1998), PBL – Problem Based Learning ou APP – Apprentissage Par Problèmes, in D. Leclercq (Ed.), Pour une pédagogie universitaire de qualité, Sprimont : Mardaga, pp. 187-205.

Leclercq , D., Simon, F., Marotte, P., Verschueren, A. et Lacaille, C. (2002). Former des étudiants de première candidature universitaire à des compétences transversales : Lesquelles

et comment ? Deuxième Congrès des chercheurs en Education de la CFWB, Louvain-La-Neuve.

Leclercq D., Rinaldi A.M. et Ernould C. (2003). Un questionnaire spectral pour l'évaluation des connaissances chez le patient diabétique, in Gagnayre et al. (Eds), L'évaluation de l'Education Thérapeutique du Patient, Paris : IPCEM.

Leclercq, D., Marotte, P., Massart, V., Simon, F., Poumay, M., Cabolet, C., Bolland, J. (2003). Deux approches contrastées pour développer les compétences transversales dans les grands groupes universitaires. XX° Colloque de l'AIPU, Sherbrooke.

Leclercq, D. (Ed) (2003). Diagnostic cognitif et métacognitif au seuil de l'université. Liège : Editions de l'université de Liège.

Leclercq, D., Detroz, P ., Dupont, C., Gilles, J.L. (2003). L'opération RESSAC, in D. Leclercq, (Ed).Un diagnostic cognitif et métacognitif au seuil de l'université. Le projet MOHICAN mené par les 9 universités de la Communauté Française Wallonie Bruxelles. Liège : Editions de l'Université de Liège,155-170.

Leclercq, D. et Detroz, P. (2003). Liens entre caractéristiques de départ et les réussites en première candidature, in Leclercq, D. (Ed) (2003). Diagnostic cognitif et métacognitif au seuil de l'université. Liège : Editions de l'université de Liège.

Leclercq, D. (2005) Edumétrie et docimologie pour praticiens chercheurs, Editions de l'université de Liège.

Leclercq, D. et Poumay, M. (2005) Métacognition. Chap. 7 de D. Leclercq. Méthodes de Formation et Théories de l'Apprentissage. Editions de l'Université de Liège.

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. Educational Measurement, 17, 28-30

Luh, C.W. (1922). The conditions of retention. Psychol. Monograph, 31, 142, 401-410.

Lune, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. Educational Measurement, 17, 24-28.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement. 21, 215-237.

Messick, S. (1988, 3° edition). Validity. In Linn R. (Ed), Educational Measurement,. 3d ed, American Council of Education , NY : Macmillan, 13-103.

Messick, S. (1975). The Standard Problem: Meaning and Values in Measurement and Evaluation, American Psychologist 30,: 955-66;

Moss, P. A. (1998). The role of consequences in validity theory. Educational Measurement, 17, 6-12.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. Educational Measurement, 17, 13-16.

Schurwirth, L.W.T. (1998). Computerized Case-based Testing : an approach to the assessment of medical problem solving. Ph.D. in Education, Maastricht : University of Maastricht.

Taleporos, E. (1998). Consequential validity: A practitioner's perspective. Educational Measurement, 17, 20-23, 34.

Van der Vleuten, C. & Wijnen, W. (1990). Problem-based learning : Perspective from the Maastricht experience, Amsterdam : Thesis.

Verhoeven, B. (2003). Progress testing. The utility of an assessment concept. Ph D Dissertation. Univ. Maastricht.

Yen, W. M. (1998). Investigating the consequential aspects of validity: Who is responsible and what should they do? Educational Measurement, 17, 5.