hep/
Haute école pédagogique du canton de Vaud
Avenue de Cour 33 (bât. C39) — CH 1014 Lausanne
www.hepl.ch

Approche qualité en évaluation des apprentissages

Jean-Luc Gilles

Professeur ordinaire à la Haute école pédagogique du canton de Vaud, Lausanne, Suisse.

Notes de cours de la partie introductive du cycle CGQTS
(cycle de Construction et de gestion qualité des tests standardisés)
à l'attention des étudiant.e.s du cours « *Evaluation des apprentissages* » du MASPE
(Master en sciences et pratiques de l'éducation – HEP Vaud - UNIL)

Ces notes sont extraites des pages 21 à 35 de la thèse présentée pour l'obtention du grade de docteur en sciences de l'éducation de Jean-Luc Gilles

Référence de la thèse complète :

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). Université de Liège, Liège, Belgique. http://hdl.handle.net/20.500.12162/824

Le besoin :
des évaluations de qualité

Sommaire

A. Problèmes liés aux examens oraux ou écrits ayant recours aux Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL)

Le sort académique (réussite ou échec) et par suite le sort professionnel de millions d'étudiants par le monde dépend d'évaluations sommatives certificatives. Pour des raisons d'équité, relayées par des raisons de prudence juridique (afin d'éviter les recours en justice), ces évaluations sont de plus en plus standardisées. On sait en effet depuis des décennies, notamment depuis les travaux de recherche du courant de docimologie critique mené par Pieron (1963), à quel point la correction des examens où les étudiants sont interrogés à l'aide de Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL), que ce soit par écrit ou oralement, souffre de nombreux biais. Ces biais expliquent en grande partie le succès rencontré par les épreuves standardisées dont le contenu et les conditions d'administration sont identiques pour tous les examinés et dont la correction peut être automatisée. Pour mieux comprendre les enjeux de la standardisation des épreuves universitaires, nous dressons ici un inventaire (non exhaustif) des principales faiblesses des examens ayant recours aux QROM/QROL.

1. Le manque de concordance intra et inter-correcteurs dans la correction des réponses ouvertes

La correction des Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL) ne peut être automatisée et est donc confiée à un ou plusieurs correcteurs. Nous allons voir que cette caractéristique entraîne des problèmes liés au manque d'accord intra-correcteur ou/et inter-correcteurs.

Travaillant entre autres sur le matériel fourni par le baccalauréat français, Pieron (1963) et ses collaborateurs ont montré que les notes obtenues par un examiné dépendaient de l'« attitude typologique » de l'examinateur lors de la correction des examens. Leclercq (1986, p.22) résume cette équation personnelle du correcteur en quatre caractéristiques principales qui pourraient être mesurées soit d'une épreuve à l'autre, d'une année à l'autre ou même au sein d'une seule épreuve : « *(1) le centrage ou la moyenne de ses notes (certains examinateurs se révèlent trop sévères et d'autres excessivement généreux) ; (2) l'ampleur de la distribution de leurs notes (dispersion exprimée par l'écart type) ; (3) la forme de la distribution de leurs notes (normale, symétrique ou non) ; (4) la constance dans le temps de ces trois paramètres* ».

Les ouvrages traitant de docimologie fourmillent d'études où sont mises en évidence les attitudes typologiques divergentes des correcteurs. Nous mentionnons ici à titre d'exemple le cas mis en évidence par Agazzi (1967, relaté par De Landsheere 1979, p. 33) pour un ensemble de branches où six correcteurs ont, chaque fois, noté les examens du baccalauréat[1] :

	Refusés par les six correcteurs	Admis par les six correcteurs	Admis par les uns et refusés par les autres
Version latine	40%	10%	50%
Composition française	21%	9%	70%
Anglais	37%	16%	47%
Mathématiques	44%	20%	36%
Philosophie	9%	10%	81%
Physique	37%	13%	50%

On voit que pour l'examen de philosophie, dans 81% des cas l'admission ou le refus d'un candidat dépendait du correcteur, la concordance entre correcteurs n'apparaît donc que dans 19% des cas. C'est en mathématique que le pourcentage de concordance est le plus élevé mais avec cependant seulement 66% de décisions de refus ou d'admissions sur lesquels les correcteurs s'accordent.

[1] Note de refus/échec : moins de 10 sur 20.

Des problèmes de discordance peuvent également s'observer chez une seule personne chargée de la correction, on parlera alors d'infidélité chez un même correcteur. L'étude de Hartog et Rhodes (1936), relatée par De Landsheere (1979, p. 39), est exemplaire : quatorze historiens furent invités à corriger une seconde fois quinze compositions, douze à 19 mois après la première correction (après que toute trace de la première correction fut effacée). De Landsheere rapporte : « *Dans 92 cas sur 210, le verdict a été différent d'une fois à l'autre. Il faut toutefois insister sur le fait que des résultats aussi pauvres sont dus au manque de directives rigoureuses précisant les aspects à considérer par les notateurs.* »

Leclercq (1986, p. 22) distingue trois catégories de biais liés à la correction avec deux effets dans chaque catégorie. Nous les résumons dans le tableau ci-dessous.

Biais dus au seul correcteur		Biais dus aux interactions prof.-élève		Biais dus aux séries de copies	
Effet de sévérité	Effet de tendance centrale	Effet de halo	Effet de stéréotypie	Effet de séquence	Effet de relativisation
Sévérité systématiquement plus élevée ou au contraire moins élevée chez certains correcteurs	*Evitement des notes extrêmes et concentration des scores au milieu de l'échelle*	*Des caractéristiques de l'étudiant influencent sa note (aspect physique, présentation, …)*	*Tendance à attribuer à un examiné les notes que celui-ci a acquise antérieurement*	*La copie qui suit une copie brillante risque d'être désavantagée, et inversement*	*Parmi toutes les copies jugées moyennes quelques mois auparavant le correcteur distinguera des faibles et des bonnes*

Ce bref aperçu des biais liés à la subjectivité des examinateurs montre que les problèmes d'infidélité lors de la correction sont énormes : les examens traditionnels offrent peu de garanties qu'un travail corrigé et classé dans la catégorie « excellent » bénéficierait de la même mention s'il était corrigé dans d'autres conditions (autres correcteurs ou quelques semaines plus tard par un même correcteur).

Signalons ici que les constats négatifs de la docimologie critique (tels que ceux liés à l'inconstance intra et inter-correcteurs que nous venons d'évoquer) ont cependant permis l'avènement d'un courant de docimologie constructive où les docimologistes ont tenté d'apporter des solutions (malheureusement trop peu mises en pratique). Citons à titre d'exemple les travaux sur les « Echelles descriptives en évaluation » de De Bal, De Landsheere et Beckers (1977). A l'aide de la technique des échelles descriptives ces chercheurs ont mis en évidence la possibilité d'améliorer la cohésion inter-correcteurs en découpant le « trait » à évaluer en plusieurs facettes et en veillant à définir clairement chaque échelon des échelles mises au point pour mesurer les aspects à évaluer.

2. Le manque de validité

Lors d'un examen ayant recours aux QROM/QROL, lorsque le nombre d'étudiants est élevé (parfois plus de 600 dans une institution comme l'Université de Liège) l'examinateur est contraint d'utiliser un nombre réduit de questions ouvertes étant donné le temps considérable que nécessite la correction des réponses fournies. Le faible nombre de questions posées entraîne dès lors le problème de l'absence d'une couverture large de tous les points importants du cours. Il est en effet difficile avec seulement quelques QROL de balayer l'ensemble du cours.

Une autre critique liée au problème de la validité et qui peut aussi être adressée à l'encontre des examens ayant recours aux QROM/QROL concerne la tendance à exiger la simple restitution de faits abordés dans le cours malgré la possibilité d'évaluer des processus mentaux plus complexes. En effet, parmi les processus mentaux qui sont évalués lors des examens, force est de constater que le plus souvent c'est la connaissance de mémoire qui est sollicitée et plus rarement la capacité d'analyser, de synthétiser, d'élaborer des jugements critiques, etc. De Landsheere (1979, p. 52) signale que dès 1911 un rapport de la Commission Consultative sur les Examens dans l'Enseignement Secondaire de Grande Bretagne déplorait que les élèves consacraient trop d'énergie à reproduire les idées des autres au lieu de développer leur propre créativité.

Si un examinateur souhaite évaluer le niveau « connaissance » chez les étudiants, il est préférable qu'il le fasse à l'aide d'épreuves standardisées ayant recours aux Questions à Choix Multiple (QCM). Le nombre de questions qu'il pourra poser sera bien plus élevé et lui permettra de couvrir une large partie de la matière enseignée. Nous verrons par la suite qu'il existe des formes de QCM plus sophistiquées que les QCM classiques et qui permettent d'évaluer plus systématiquement des processus mentaux plus élevés (dans la taxonomie de Bloom, 1969) que la simple connaissance de mémoire (en fait de la 'reconnaissance' dans le cas des QCM où l'examiné doit choisir, donc reconnaître, parmi *x* solutions qui lui sont proposées, celle qui est correcte).

3. Le manque de sensibilité des mesures qui ignorent les états de connaissances partielles

Tout le monde s'accorde sur le fait que les mesures des acquis des étudiants devraient refléter des phénomènes subtils mais bien peu d'examinateurs se soucient de sonder méthodiquement la conviction avec laquelle les examinés maîtrisent le sujet sur lequel ils sont interrogés.

Rarement on demande aux étudiants d'exprimer systématiquement leurs certitudes à propos des réponses qu'ils fournissent lors d'un examen. C'est ignorer qu'en termes de connaissances partielles, un fossé peut séparer la performance de deux étudiants qui fournissent pourtant la réponse correcte à une même question. En effet, la performance de l'examiné qui répond correctement et de façon très assurée (en accompagnant sa réponse d'une probabilité élevée d'être correcte) est bien meilleure que celle d'un autre examiné qui lui aussi fournit la réponse correcte mais en lui attribuant une probabilité si peu élevée de l'être que cette (pseudo) connaissance en devient inutilisable parce que le sujet lui-même ne peut se fonder sur elle pour prendre des décisions, pour agir. De même, les étudiants qui avouent leur ignorance devraient être moins sanctionnés (et même encouragés dans cette démarche) par rapport à ceux qui prétendent avec assurance fournir des réponses correctes qui s'avèrent erronées (ce qu'Ebel a appelé « *unwarranted pretense of confidence* »). Soulignons que dans ce dernier cas les « prétentions de connaissances erronées » constituent des comportements particulièrement dangereux dans les domaines où la vie des gens est en jeu (par exemple en médecine, en pilotage d'avion, de véhicules,…).

Malheureusement, les examinateurs se donnent trop peu souvent les moyens d'évaluer les connaissances partielles des étudiants, et, lorsqu'ils le font dans le cadre d'examens traditionnels c'est en général de façon fort peu systématique. Signalons qu'on commence à en comprendre l'intérêt dans l'industrie (Shufford, 1993) et dans l'éducation du patient (D'Ivernois et Gagnayre, 1995).

Or, la recherche et la pratique ont montré qu'en respectant une série de règles méthodologiques (De Finetti, 1965; Shufford & al., 1966; Van Naerssen, 1962; Leclercq, 1975, 1982, 1993; Hunt, 1977; Bruno, 1986, 1987) il est possible de mesurer de manière subtile, systématique et objective les états de connaissances partielles des étudiants à l'aide des pourcentages de certitude. Lorsqu'ils sont utilisés, ceux-ci permettent à l'examiné d'exprimer son doute en accompagnant chacune de ses réponses de la probabilité qu'il lui accorde d'être correcte. Nous passons alors d'une conception binaire et frustre de la mesure des compétences à une conception spectrale et subtile où il est enfin possible de distinguer selon la terminologie proposée par Jans & Leclercq (1999, p. 307) entre (de la situation la plus catastrophique à l'idéal) : (1) méconnaissance erronée (réponse incorrecte et certitude élevée), (2) confusion (réponse incorrecte et certitude moyenne), (3) méconnaissance reconnue (réponse incorrecte et certitude faible), (4) ignorance (réponse correcte et certitude zéro), (5) connaissance douteuse (réponse correcte et certitude faible), (6) connaissance partielle (réponse correcte et certitude moyenne) et (7) connaissance parfaite (réponse correcte et certitude élevée).

Nous reviendrons sur les techniques de recueil des pourcentages de certitude qui sont à la base des indices de qualité spectrale des épreuves.

Bien qu'il soit possible de demander aux étudiants d'accompagner leurs réponses ouvertes moyennes ou longues par des pourcentages de certitude, cette procédure n'est que très rarement utilisée lors

d'exercices et quasi jamais lors d'examens où le professeur sollicite des réponses ouvertes[2]. C'est dans le cadre d'épreuves standardisées ayant recours aux QCM avec pourcentages de certitude que l'évaluation systématique des états de connaissances partielles est la plus aisément mise en œuvre.

4. Le manque de diagnosticité des épreuves sommatives classiques qui ont recours aux QROM ou aux QROL

Les examinateurs se préoccupent en général très peu de renvoyer après l'épreuve un feedback détaillé et individualisé vers les examinés de manière à permettre à ces derniers d'effectuer un bilan précis de leurs compétences. Certains prétexteront que les examens certificatifs sont là pour vérifier si l'étudiant est capable de réaliser les tâches qu'on attend de lui en fin d'enseignement et non pour diagnostiquer où se situent les éventuels problèmes qu'il rencontre, qu'il ne faut pas confondre évaluation formative à visée diagnostique et évaluation sommative à visée certificative.

Cependant, dans notre contexte universitaire les étudiants qui subissent un échec dans un cours en 1ère session disposent en général d'une deuxième chance sous la forme d'une seconde épreuve en 2ème session (souvent du même type que la première) et de quelques mois pour améliorer leurs connaissances. Dès lors, pourquoi ne pas renvoyer un maximum d'informations vers ces étudiants en situation d'échec, et, le plus tôt possible après l'épreuve, de manière à leur permettre d'ajuster leur étude en fonction des faiblesses décelées lors du premier examen ? Ne pas le faire est à notre avis condamnable et relève de la « *non assistance à étudiant en danger d'échec* ».

Il existe des explications moins avouables à ce manque de rétroaction vers les examinés (explications d'autant plus difficiles à comprendre qu'un des chevaux de bataille de la plupart des institutions universitaires est la lutte contre l'échec). Ainsi, on avancera le fait que les épreuves sommatives traditionnelles ayant recours aux QROM/QROL ne sont en général pas conçues pour permettre des feedbacks détaillés. Il est vrai que quelques questions ouvertes ne couvrant qu'une partie très limitée de la matière (voir le manque de validité des épreuves traditionnelles) ne permettent guère d'informer l'étudiant sur ce qui est maîtrisé ou non dans son étude du cours.

La vérité est que les examinateurs évitent de poser de nombreuses questions à réponses ouvertes notamment parce que la correction de celles-ci prend un temps considérable. Un des avantages des épreuves standardisées ayant recours à des questions fermées est de permettre de poser de nombreuses questions couvrant une large partie de la matière et dont la correction peut être automatisée (donc effectuée en peu de temps). Nous verrons plus loin que cette automatisation de la correction des épreuves standardisées permet aussi d'envisager l'envoi de feedbacks détaillés (notamment via Internet) vers les étudiants, et ce, dans des délais très courts.

5. Le manque d'équité des épreuves traditionnelles, en particulier les oraux

En période d'examen on entend souvent, dans les couloirs jouxtant les bureaux des examinateurs, des commentaires de la part des étudiants sur la chance ou la malchance qui fut la leur en ce qui concerne le tirage au sort des questions. Il faut reconnaître que le facteur chance peut en effet jouer un rôle important dans la réussite lorsque seulement deux ou trois questions sont posées lors d'une épreuve.

Les étudiants savent aussi que lors des examens oraux il vaut mieux ne pas suivre un condisciple particulièrement brillant car l'effet de contraste risque d'être défavorable. Inversement, il vaut mieux suivre un condisciple qui a échoué pour autant que la contre-performance de ce dernier n'ait pas mis le professeur de trop mauvaise humeur… La chance peut donc aussi jouer au niveau de l'ordre de passage chez l'examinateur.

Ce type de phénomène a été observé par Bonniol (1972) dans le cadre de la correction de copies d'épreuves écrites. Ce dernier introduit dans la correction de travaux de valeur moyenne des ancres (copies

[2] D. Leclercq l'a fait dans une épreuve de vocabulaire français en novembre 2000 avec des étudiants de 1ère candidature à la FAPSE-ULg et Jans (1997) dans le domaine du vocabulaire anglais.

de valeur soit excellente, soit médiocre) et constate un effet de contraste sur la note attribuée aux travaux suivants : une copie de qualité moyenne sera surévaluée après une copie médiocre ou sous-évaluée après une copie excellente.

Ces quelques exemples et les problèmes liés au manque de concordance intra-correcteur et/ou inter-correcteur exposés plus haut (p. 5) montrent que les examinés peuvent se retrouver dans des situations fort injustes parce qu'elles manquent de standardisation (tous ne sont pas traités de la même façon) et parce qu'une trop grande place est laissée au facteur chance.

Il existe un sentiment d'injustice chez les étudiants lié à ce manque d'équité des examens traditionnels. Comme le soulignent les auteurs de l'enquête sur les pratiques d'examen à l'université de Montréal (Blais & al., 1997, pp.126-127) : « *...lors des entrevues de groupe avec les étudiants, la plainte la plus fréquente est celle qui concerne le manque de standardisation de l'évaluation et de la notation (...). Les étudiants en ont contre le fait que les professeurs les évaluent comme bon leur semble. Ils dénoncent la subjectivité qui intervient dans l'élaboration et la correction des travaux ou examens. Pourquoi dans certains domaines les étudiants trouvent-ils facile de « passer » et d'obtenir des notes élevées (on pourrait parler d'inflation de notes), alors que dans d'autres domaines, les étudiants sont continuellement sous le stress d'une évaluation exigeante leur demandant, selon leur perception, de fournir des efforts plus grands que leurs collègues ? Que dire des programmes où il existe des cours « éliminatoires », plus difficiles que les autres et faisant office de mécanisme de sélection ? De leur point de vue, il s'agit d'injustices flagrantes, du point de vue de l'enseignant, il s'agit de choix « personnels » en relation avec l'absence de balises contraignantes (voir institutionnelles) quant aux exigences de réussite des cours* ».

Après cette revue rapide des biais potentiels liés aux examens traditionnels, on comprendra probablement mieux l'ironie et l'humour provocateurs qui caractérisent les propos des auteurs du livre intitulé « *Le bouton du mandarin* », Didier & al. (1966), lorsqu'ils définissent le sens primitif et le sens dérivé du terme « examen » :

Sens primitif	Sens dérivé
« *Epreuve, ayant pour objet de vérifier l'état physique, intellectuel ou moral, de la chose ou de la personne examinée* »	« *Loterie instituée pour la distribution de lots appelés 'diplômes'* »

La situation est cependant pire, le hasard dans les examens, surtout dans les oraux, où les QROM/QROL sont utilisées, intervient moins que ne le laisse penser le terme « *loterie* » employé par les auteurs.

En effet, les épreuves traditionnelles ne sont pas socialement neutres. On sait que les examinés selon qu'ils proviennent de classes sociales aisées ou défavorisées n'emballeront pas leurs compétences de la même façon. Ce phénomène a été bien mis en évidence par Passeron (1970). Ce dernier (p. 12) propose la thèse de l'examen instrument d'immobilisme social : « *Les procédures de notation et les types d'épreuves utilisées prennent en compte au moins autant que les aptitudes techniques certains aspects gratuits de la performance, qui n'ont aucune importance technique, mais qui sont en revanche très fortement liés aux habitudes culturelles de telle classe sociale plutôt que telle autre* ». Tourneur (1988, p. 4) commente le texte de Passeron en signalant : « *La critique vaut surtout pour l'examen oral si prisé dans l'enseignement supérieur et qui accorde la part belle à la présentation, à la correction du langage et à l'élégance de la diction* ».

Avec l'accentuation de la diversification des origines sociales des étudiants universitaires, les examinés sont de plus en plus conscients des problèmes liés à l'effet de halo lors des oraux. Voici l'extrait d'une interview d'un étudiant en pleine session d'examen parue dans le quotidien Le Soir[3] du 29 mai 1999 : « Daniel, en deuxième licence philo et lettres, s'inquiète pour des raisons économiques. *Je ne peux pas rater: c'est le CPAS qui m'a pris en charge. C'est très paniquant, surtout maintenant que je suis si proche du but. J'ai très peur des "oraux", parce que je n'arrive pas à m'exprimer facilement. J'ai grandi dans la rue avec beaucoup d'immigrés autour de moi. On a toujours parlé une langue qui n'a rien à voir*

[3] Article intitulé « *Tremblez maintenant !* » de Michel Verlinden.

avec celle qu'on pratique ici. Devant un prof, j'ai peur de ce que je dis, surtout que certains ne se gênent pas pour vous faire remarquer vos points faibles... ».

Même si nous pensons que le déterminisme social n'est pas inéluctable, il faut cependant bien admettre que l'effet de halo ne joue pas en faveur des étudiants provenant de couches sociales moins favorisées. Si dans un pays comme les Etats-Unis les tests standardisés connaissent un tel succès, c'est notamment parce qu'ils offrent des garanties de non discrimination raciale (*culture-free test*) et de non discrimination de classe sociale (*class-free test*).

***

En synthèse :

A la lecture des problèmes posés par les épreuves traditionnelles ayant recours aux Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL) on aura compris que l'enjeu principal des épreuves standardisées est d'éviter que la réussite ou l'échec des étudiants soit tributaire de la mauvaise qualité des procédures et des instruments d'évaluation qui font la part belle à la subjectivité des examinateurs. Faut-il s'étonner que les USA nous précèdent depuis des décennies dans cette démarche de standardisation, eux qui sont obnubilés par la problématique de la non discrimination et de l'objectivité des processus de sélection ?

Malgré les biais bien connus des QROM/QROL et les difficultés liées à leur correction lorsque les professeurs doivent faire face à de grands groupes avec peu de moyens, nous ne préconiserons pas l'abandon de cette modalité de questionnement mais son amélioration et son usage en complémentarité avec les QCM standardisés. La procédure des « échelles descriptives » (De Bal, De Landsheere et Beckers, 1977) permet d'améliorer la fidélité de la correction des QROM/QROL qui demeurent indispensables pour mesurer certains types de performances complexes.

B. Les évaluations standardisées permettent-elles de faire mieux ?

L'exposé des points faibles des examens traditionnels ayant recours aux Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL) pourrait amener certains à conclure qu'il faut définitivement abandonner cette forme de questionnement et s'orienter exclusivement vers des épreuves standardisées ayant recours aux questions fermées. Nous ne le pensons pas. Nous commencerons par montrer dans cette section que les examens standardisés présentent eux aussi des faiblesses liées à l'utilisation des Questions à Choix Multiples (QCM) classiques[4]. Il existe des parades à ces problèmes, elles seront brièvement évoquées ici et décrites plus en détail dans le chapitre suivant. Plutôt que d'opposer les examens traditionnels à réponses ouvertes aux examens standardisés ayant recours aux QCM nous préconisons une complémentarité entre ces deux types d'approches lorsque les objectifs de l'évaluateur l'exigent. Nous terminerons en insistant sur le fait qu'il existe des épreuves standardisées dont les niveaux de qualité peuvent être très différents. Nous verrons à ce propos qu'on peut opposer une approche « amateuriste » à une autre plus « professionnelle » pour la construction des épreuves standardisées universitaires.

1. Les examens standardisés classiques sont sensibles à une série d'inconvénients inhérents aux questions à choix multiple classiques

Bien que l'automatisation de la correction permette d'échapper aux incohérences intra-correcteur ou/et inter-correcteurs, les examens standardisés ont aussi leurs inconvénients. Ces problèmes sont inhérents aux questions à choix multiple. Nous évoquerons ici cinq problèmes qui nous paraissent devoir être signalés mais qui ont tous leur parade.

a) Le piège de la parcellisation des connaissances

Les concepteurs de QCM sont souvent tentés de poser des questions de détail car plus la réponse attendue est précise et limitée à un contexte très particulier et moins elle sera susceptible d'être contestée. Les QCM doivent être « décidables », elles ne peuvent pas être sujettes à interprétation. Noizet et Caverni (1978) cités par Leclercq (1986, p. 31) signalent à ce propos : « *Cette condition de décidabilité est capitale puisque la réponse sollicitée est de nature 'vrai' ou 'faux'. Ainsi, par exemple, la question de savoir si tel écrivain a écrit ou non telle œuvre constitue un item décidable dans la mesure où la paternité d'une œuvre est susceptible d'être attribuée sans erreur. Mais la question de savoir si en dernière analyse tel écrivain peut être considéré comme un romantique constituerait un item indécidable* ».

Cette caractéristique entraîne une tendance à poser des questions de détails dans les examens standardisés. Les effets de cette tendance se manifestent alors aussi dans la façon dont les étudiants abordent l'étude du cours : ils se focalisent sur les détails en ne prenant par exemple plus la peine d'effectuer les liens entre les différentes parties du cours. Des solutions existent pour contrer cette *tendance à la parcellisation des connaissances*.

D'une part, Leclercq (1993) a montré qu'on peut améliorer les QCM à l'aide des Solutions Générales Implicites (SGI), ce qui permet à l'évaluateur d'éviter de ne poser que des questions de connaissance. Mais cette sophistication des consignes entraîne alors aussi d'autres difficultés (surmontables) liées à la nécessité d'un entraînement préalable des examinés. Nous verrons plus loin quels dispositifs nous mettons en place pour permettre aux étudiants de s'entraîner avant l'examen final.

D'autre part, un accompagnement méthodologique des examinateurs par des personnes expérimentées dans le domaine de la création des QCM, diminue le risque de tomber dans le piège de la rédaction des questions de détails. Des procédures de contrôle de la qualité des questions existent et

[4] Nous ajoutons l'appellation « classique » à QCM de façon à les différencier des Questions à Choix Multiple avec Solutions Générales Implicites (QCM-SGI) (Leclercq, 1993) qui permettent de palier à une série d'inconvénients que présentent les QCM *classiques*.

permettent d'améliorer la qualité des épreuves (qualité *a priori*). Il existe des possibilités d'analyse de la qualité *a priori* des QCM, notamment par la relecture formelle des questions en fonction d'une série de règles méthodologiques décrites par Leclercq (1986, pp. 79-144).

b) Le danger de la mémorisation des réponses incorrectes aux questions fermées

On peut craindre que les étudiants mémorisent les solutions incorrectes des QCM lors des examens standardisés. Cette crainte est ancienne et avait déjà été exprimée il y a une quarantaine d'années par Skinner (1961). Elle fut ensuite confirmée par Preston (1965) mais lors de son expérience, ce dernier n'avait pas fourni les solutions correctes aux sujets testés après l'épreuve.

C'est Karraker (1967) qui a démontré que ce danger n'est plus à craindre lorsqu'on communique les réponses correctes et que, au contraire, le testing via des QCM améliore les performances à une épreuve ultérieure ayant recours à des questions ouvertes construites sur la même matière. Les expériences de Preston et de Karraker liées au problème de la mémorisation des solutions fausses ont été décrites en détail par Leclercq (1986, pp. 35-40).

La parade de ce danger de mémorisation des réponses incorrectes se situe donc dans la rétroaction qui doit suivre rapidement la passation de l'examen. Nous verrons plus loin que le principal reproche que nous adressent les étudiants lorsqu'on leur demande leur avis sur les examens, se situe justement au niveau du manque de feedback. Des solutions technologiques existent pour permettre une diffusion rapide et détaillée vers les étudiants et via l'Internet des résultats d'un examen.

c) Le manque d'équité lorsque les possibilités de fraudes ne sont pas suffisamment prises au sérieux

La nature des réponses fournies aux examens standardisés ayant recours aux QCM : une lettre ou un chiffre désignant la réponse choisie, permet de poser beaucoup de questions et de simplifier la correction en l'automatisant. Malheureusement, le revers de la médaille réside dans le fait que ce type de réponses peut être assez facilement communiqué. Il suffit par exemple que les tricheurs conviennent avant l'épreuve d'un code gestuel différent pour chaque proposition d'une QCM. La communication des réponses devient alors un jeu d'enfant lors de la passation de l'épreuve. L'épreuve n'est dès lors plus équitable dans la mesure où les examinés qui ne cèderont pas à la tricherie seront pénalisés.

Des parades existent. Elles consistent à fournir des versions parallèles d'un même questionnaire. Les étudiants restent soumis aux mêmes questions, mais celles-ci sont mélangées au sein des différentes versions. Le danger d'épreuves standardisées biaisées par des fraudes est aussi lié à des conditions d'administration inadéquates. Les locaux où ont lieu ces épreuves doivent être adaptés à la taille des groupes d'étudiants évalués, assez grands que pour permettre d'espacer suffisamment les examinés. Le nombre de surveillants et le sérieux avec lequel ces derniers accomplissent leur tâche constituent aussi des facteurs à prendre en compte pour éviter la fraude.

En fait, l'organisation d'un examen, particulièrement lorsqu'on est confronté à un grand groupe d'examinés, demande un effort de préparation intense où de multiples facteurs doivent être passés en revue dont les mesures anti-fraude. En guise d'exemple à ne pas suivre, on se rappellera des évènements qui eurent lieu lors d'un grand concours de recrutement de la Communauté européenne en septembre 1998. Suite à ce concours, le quotidien Le Soir du 21 septembre 1998 titrait : « *Tricheries et contestations au concours de recrutement - La Commission européenne a raté son examen* ». Voici un extrait de l'article[5] « *…Tous les témoignages concordent. Au Heysel, beaucoup ont pu tricher sans encourir la moindre sanction (…) on s'échangeait des réponses, on téléphonait carrément à l'extérieur, à l'aide d'appareils portables, pour obtenir des renseignements. (…) Dans un des centres d'examen à Rome, la situation était tellement confuse … qu'un candidat a appelé la police pour venir constater les irrégularités. Ce qui fut fait…* ». L'absence de procédures anti-fraude sérieuses et le manque de contrôle de la situation par les

[5] Article signé par André Riche.

organisateurs de ce concours qui réunissait plus de 30.000 candidats à travers les quatre coins de l'Europe provoqua l'annulation de l'épreuve et la Commission européenne fut amenée à produire des excuses publiques.

Si l'on calcule en heures de travail perdues dans la préparation par les organisateurs et surtout dans la passation par les candidats, la perte est gigantesque sur le plan matériel. Sans parler des dommages moraux (pour les individus) et de l'image de l'Union Européenne.

d) Le problème des réponses aux hasard

Les solutions les plus anciennes au problème des choix au hasard dans les QCM ont consisté à augmenter le nombre de propositions ou/et à pénaliser les erreurs en appliquant la méthode de la *correction for guessing*. Cependant ces deux solutions ont leurs inconvénients. En ce qui concerne l'augmentation du nombre de propositions, il est bien connu que certaines QCM n'ont « naturellement » que peu de solutions[6] et en ajouter revient à créer une proposition dont l'attractivité sera très faible. Pour ce qui est des pénalités en cas de réponses au hasard, les travaux de Leclercq, (1987) ont montré que la *correction for guessing* est d'autant plus un instrument à bannir, qu'il peut être remplacé par l'utilisation des probabilités subjectives qui, elles, (1) sont basées sur un modèle théorique plus pertinent, (2) ont un principe de notation plus équitable, (3) sont plus formatives pour l'apprenant et (4) plus informatives pour l'examiné et l'enseignant.

e) Les QCM ne permettent pas de mesurer tous les types de performances

Cette critique est souvent formulée à l'encontre des QCM et il est vrai que ce type de questionnement ne permet pas actuellement de mesurer la capacité à rédiger, à inventer, à s'exprimer oralement... A ce propos, Leclercq (1986, p. 34) rappelle : « *Les QCM sont un outil parmi d'autres : il importe de recourir au mode d'évaluation le plus adéquat à chaque situation. Les QCM conviennent moins bien pour les performances complexes (réponse longue où la structure et l'expression jouent un grand rôle) que pour les performances isolables* ».

Si l'évaluation de certains types de performances ne peut en effet s'effectuer qu'à l'aide de réponses ouvertes, il faut associer cette forme de questionnement aux examens standardisés quand les objectifs de ces derniers l'exigent. C'est la solution que nous proposons aux enseignants.

On pourrait également craindre qu'avec la généralisation des examens standardisés et l'impossibilité d'exprimer une réponse longue à l'aide des QCM, les étudiants soient de moins en moins sollicités à s'exprimer par écrit ou oralement et donc « perdent » cette compétence. Ceci constitue à notre avis un argument supplémentaire en faveur de formules d'évaluation faisant appel à des combinaisons QCM-QROM/QROL.

[6] Par exemple, le genre d'un nom commun ne peut être que féminin, masculin ou neutre, dès lors, inutile d'inventer une quatrième solution étant donné que les propositions sont « imposées » par la nature du contenu de la question.

2. La nécessité d'une complémentarité entre QCM (de qualité) et QROM/QROL (améliorées) lorsque les objectifs de l'évaluation l'exigent

L'exposé des principaux biais liés aux Questions à Réponses Ouvertes Moyennes (QROM) ou Longues (QROL) pourrait amener le lecteur à croire que ce type de questionnement est à proscrire. Mais comme nous venons de le voir, les QROM/QROL demeurent indispensables pour évaluer certains types de performances complexes : capacité à synthétiser, développement de raisonnements, créativité, aptitude à s'exprimer par écrit, etc. qui ne peuvent pas être mesurées à l'aide de questions fermés.

Malheureusement, la façon dont les QROM/QROL sont habituellement corrigées fait la part belle à la subjectivité des examinateurs. Cependant, la fidélité des correcteurs peut être améliorée, De Bal, De Landsheere et Beckers (1977) ont en effet montré à l'aide des « *Echelles descriptives en évaluation* » qu'il était possible d'améliorer la cohésion inter-correcteurs en découpant le « trait » à évaluer en plusieurs facettes et en veillant à définir clairement chaque échelon des échelles mises au point pour mesurer les aspects à évaluer. Cependant, c'est un autre inconvénient qui apparaît alors : la tâche demandée à l'évaluateur est encore plus complexe vu l'effort d'analyse nécessaire pour séparer les différentes catégories de réponses. Cette technique des échelles descriptives augmente la fidélité de la correction mais ne diminue pas le temps de correction. Ceci ne favorisera pas l'augmentation du nombre de QROM/QROL dans les épreuves lorsque de grands groupes d'étudiants sont évalués. C'est dommage car plus de QROM/QROL lors des examens permettrait de couvrir un plus large éventail de matières enseignées, donc d'améliorer la validité de contenu liée à ce type de questionnement.

D'autres techniques permettant d'augmenter la fidélité de la correction des QROM/QROL existent. Par exemple celle qui implique une double correction et qui consiste à faire corriger les copies indépendamment par deux correcteurs (A et B) qui se sont préalablement concertés sur les critères de correction. Après avoir corrigé les copies, les correcteurs confrontent question par question chaque paire de note (celle du correcteur A avec celle du correcteur B), lorsque les résultats coïncident la note est confirmée. Lorsque deux résultats diffèrent, les correcteurs se concertent en vue de comprendre ce qui a fait les différences d'appréciation. Malheureusement, cette méthode est coûteuse en temps et en personnes.

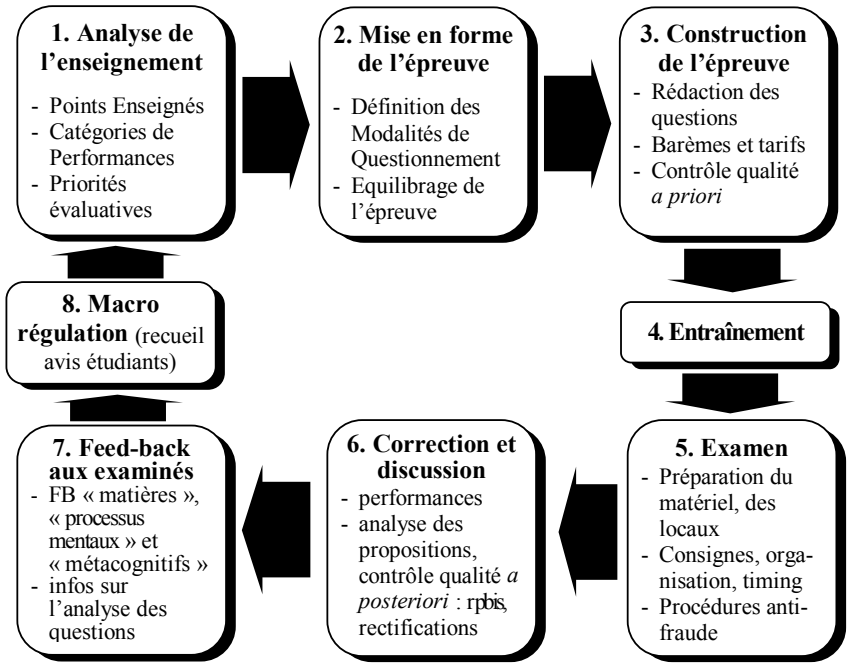
En fait, les questions à réponses ouvertes moyennes ou longues écrites améliorées à l'aide de techniques telles que les échelles descriptives, devraient être utilisées en complémentarité avec les épreuves standardisées ayant recours aux QCM. C'est ce que nous préconisons dans le cadre de l'aide méthodologique que nous proposons aux enseignants. Dans ce contexte, nous sommes souvent amené à conseiller aux examinateurs lorsque leurs objectifs évaluatifs l'imposent, d'utiliser une ou deux QROM/QROL en combinaison avec de nombreuses QCM (au minimum une trentaine). Ainsi, des QROM/QROL corrigées avec la technique des échelles descriptives permettent de mesurer des performances complexes du type de celles que nous avons évoquées plus haut tandis que le nombre élevé de QCM permet une couverture large de la matière. Une partie du gain de temps considérable lié à la correction automatisée des QCM peut ainsi être réinvestie dans la correction plus poussée d'une ou deux QROM/QROL avec échelles descriptives.

A l'aide de QCM avec Solutions Générales Implicites (QCM-SGI) nous pouvons évaluer d'autres niveaux de processus mentaux qui ne peuvent l'être, ou alors moins systématiquement, avec les QCM classiques.

3. Approche « amateuriste » et approche « professionnelle » dans la réalisation des examens standardisés

Nous pensons qu'il est nécessaire d'examiner de plus près huit étapes dans la réalisation des examens standardisés universitaires (Gilles et Leclercq, 1995). Ces huit étapes font partie d'un processus cyclique en « spirale de qualité » qui sera décrit en détail au chapitre suivant.

Les huit étapes sont résumées dans les encadrés du schéma ci-contre.



Selon nous, la construction d'un examen standardisé de qualité doit passer par ces huit étapes. Tout au long de la réalisation, le(s) concepteur(s) de l'évaluation devront veiller à rencontrer dans les procédures qu'ils mettent en œuvre, une série d'objectifs « qualité ».

D'une part, un examen standardisé ne pourra offrir des garanties sérieuses en ce qui concerne l'atteinte des objectifs « qualité » que nous listons ci-dessous qu'à la condition que la construction de l'épreuve soit envisagée dans le cadre d'un processus global comportant une série d'étapes dont les procédures et les dispositifs d'ingénierie docimologiques sont étudiés pour rencontrer ces objectifs « qualité ». Il est important de prévoir, au sein des procédures utilisées dans les phases de construction de l'examen, des micro régulations qui permettent d'améliorer la qualité du produit/service fourni.

D'autre part, nous pensons que la qualité des examens peut être améliorée d'épreuve en épreuve notamment à l'aide des avis des étudiants. Cette amélioration de la qualité des tests passe aussi par le perfectionnement du processus qui a permis de les construire, c'est l'idée du cycle en « spirale de qualité » de réalisation des examens standardisés. A ce point de vue, la dernière étape « 8. Macro régulation » est cruciale car elle permet de récolter les remarques des examinés en vue d'améliorer la réalisation des examens suivants.

Voici nos **objectifs « qualité »**. Lors de la construction d'un examen standardisé universitaire de qualité il s'agit d'offrir des garanties en ce qui concerne

- la ***validité*** : les scores des étudiants doivent refléter ce que l'enseignant veut mesurer ;
- la ***fidélité*** : un travail corrigé et classé dans une catégorie donnée doit bénéficier de la même mention s'il est corrigé dans d'autres conditions, par exemple par d'autres correcteurs ou quelques semaines plus tard ;
- la ***sensibilité*** : la mesure doit être précise ;
- la ***diagnosticité*** : le diagnostic précis des difficultés d'apprentissage, des processus maîtrisés et de ceux qui ne le sont pas doit être possible ;
- la ***praticabilité*** : la faisabilité en termes de temps, de ressources en personnel et en matériel doit est assurée ;

- l'***équité*** : tous les étudiants doivent être traités de façon juste, en principe de la même manière (standardisation) ;
- la ***communicabilité*** : les informations non confidentielles relatives au déroulement du processus doivent être communiquées et comprises par les partenaires, enseignants, étudiants, équipe de soutien docimologique engagés dans la réalisation des épreuves.

On peut dès lors opposer deux types d'approches. La première, « professionnelle », a l'ambition d'envisager le problème de la réalisation des examens universitaires de façon globale et avec la volonté de garantir la qualité et le sérieux auxquels les étudiants ont droit. C'est en effet leur sort académique qui est en jeu dans ces épreuves. La seconde, que nous qualifierons (sans doute avec excès) d'« amateuriste » consiste à négliger ou à ne pas prendre en compte une série d'objectifs « qualité » et à faire l'impasse sur des étapes pourtant cruciales dans la construction d'un examen. Cette dernière approche est souvent liée à un contexte de surcharge de travail chez certains enseignants universitaires. Dans ce cas, beaucoup préfèrent alors mettre la priorité sur les tâches de recherche plutôt que sur celles liées à l'enseignement. Il faut dire que le corps académique était, jusqu'il y a peu, essentiellement évalué sur le nombre et la qualité de ses publications dans les revues scientifiques et très peu sur la qualité des enseignements dispensés.

Depuis quelques années, en Communauté Française de Belgique (CFB), le Conseil des Recteurs des institutions universitaires Francophones (CReF) a confié à un groupe de travail la mission de mettre en place une évaluation de la qualité de ses neuf institutions universitaires (Boucher & al. 1997). Dans les rapports d'auto-évaluation et plus tard d'évaluation externe, les institutions doivent et devront aussi répondre à des questions qui envisagent la qualité des enseignements. La qualité des examens universitaires dispensés au sein d'une institution pèsera probablement encore plus qu'aujourd'hui dans ces évaluations.

On voit bien tout le bénéfice que peuvent retirer les professeurs (et par-delà l'institution universitaire de plus en plus amenée à devoir fournir la preuve de la qualité de ses enseignements) d'une structure d'appui méthodologique et logistique d'aide à la réalisation d'examens standardisés comme celle qui fonctionne actuellement à l'Université de Liège. Signalons ici une série de services qui s'insèrent dans la perspective du cycle en « spirale de qualité » et qui tiennent compte des objectifs « qualité » décrits plus haut :

- le choix du(des) type(s) de questionnements, de la(des) méthode(s) de test la(les) plus appropriée(s) ;
- la gestion des banques de questions ;
- l'analyse de la qualité *a priori* des épreuves (tests formatifs ou examens) ;
- l'entraînement des étudiants aux procédures d'évaluation ;
- la préparation et la logistique des épreuves ;
- la correction des tests à l'aide de procédures informatisées ;
- l'analyse de la qualité *a posteriori* des questions ;
- la mise en œuvre de solutions en vue de rectifier les épreuves lorsque des problèmes sont détectés ;
- la réalisation et le renvoi vers les étudiants des feedbacks individualisés relatifs à leurs performances et à la qualité de l'épreuve ;
- le recueil et l'analyse des avis des étudiants sur la qualité des épreuves.

***

En synthèse :

Les examens standardisés présentent une série de points faibles. Ceux-ci sont liés aux problèmes (bien connus) que peuvent présenter les QCM classiques. Nous verrons plus loin que les QCM-SGI et les pourcentages de certitude permettent d'améliorer notablement la qualité des épreuves ayant recours aux QCM.

Les examens standardisés offrent potentiellement de nombreux avantages : étudiants tous traités de la même façon (équité), correction automatisée (fidélité), large éventail de la matière évalué (validité de contenu), systématisation des pourcentages de certitude (sensibilité), rétroactions rapides à l'aide de feedbacks détaillés automatisés (diagnosticité et communicabilité), …

Malgré tous ces avantages, il n'en reste pas moins vrai que certaines performances complexes ne peuvent être évaluées à l'aide des QCM (par exemple la capacité à s'exprimer par écrit). Dès lors, lorsque les objectifs de l'évaluateur l'exigent, nous préconisons une utilisation combinée de l'approche standardisée et de l'approche plus traditionnelle d'évaluation à l'aide de questions ouvertes (améliorées notamment à l'aide des « Echelles descriptives en évaluation » de De Bal, De Landsheere et Beckers, 1977).

Ceci étant, les avantages potentiels liés aux épreuves standardisées et à une éventuelle approche combinée avec les questions ouvertes ne constituent pas en soi des garanties automatiques de qualité. Nous pensons que cette qualité ne peut être obtenue que dans le contexte d'une approche « professionnelle » de la réalisation des examens universitaires. Cette approche est selon nous sous-tendue par deux principes de base : (1) la qualité d'un examen standardisé universitaire ne peut être établie qu'à la condition que sept objectifs « qualité » soient atteints : validité, fidélité, sensibilité, diagnosticité, praticabilité, équité, communicabilité et (2) pour atteindre ces objectifs « qualité » il faut concevoir la construction de l'épreuve à l'aide d'un processus où les régulations permettent non seulement l'amélioration de l'examen en cours de construction, mais aussi l'amélioration du processus de construction lui-même (« spirale de qualité »*Erreur ! Signet non défini.*).

Dès lors, la réponse à la question que nous nous étions posée dans le titre de cette section « Les examens standardisés permettent-ils de faire mieux ? » peut se résumer, à ce stade, de la façon suivante.

Les épreuves standardisées permettent d'évaluer un large spectre de performances, mais pas toutes. Dans certaines situations une approche combinée incluant des questions à « réponse ouverte » dont la fidélité de la correction sera améliorée doit être recommandée. Dans tous les cas, pour faire mieux, les objectifs « qualité » entrevus ci-dessus devront être atteints, ce qui implique de concevoir la construction des examens à l'aide d'un processus méthodique dont les régulations permettent une amélioration continue des épreuves. L'analyse de la qualité a posteriori des questions est effectuée lors de la sixième étape « correction et discussion » de ce processus.

Bibliographie

Agazzi, A. (1967). *Les aspects pédagogiques des examens*, Strasbourg : Conseil de l'Europe,C.C.C.

Blais, J.-G., Laurier, M., Van der Maren, J.-M., Gervais, C., Lévesque & M., Pelletier, G. (1997). *L'évaluation des apprentissages à l'Université de Montréal et dans ses écoles affiliées*. Montréal : Université de Montréal, Faculté des sciences de l'éducation, Groupe de Recherche Interdisciplinaire en Pédagogie Universitaire (GRIPU).

Bloom, B. (1969). *Taxonomie des objectifs pédagogiques - I. Domaine cognitif*, traduit par M. Lavallée. Montréal : Education Nouvelle.

Bloom, B., Engelhart, M.-D., Forst, E.-J., Hill, W.-H. & Krathwohl, D.-R. (1956). *Taxonomy of educational objectives : handbook I, cognitive domaine*. New York : D. Mac Kay.

Bonniol, J.-J. (1972). *Les comportements d'estimation dans une tâche d'évaluation d'épreuves scolaires - Etude de quelques-uns de leurs déterminants*. Aix-En-Provence : Université de Provence.

Bruno, J. (1986). Assessing the Knowledge base of students : An information theoric approach to testing. *Measurement and evaluation in counseling and development*, 19(3), 116-130.

Bruno, J. (1987). Admissible probability measurement ininstructional management. Journal of computer based instruction, 14(1), 23-30.

Castaigne, J.-L., Gilles, J.-L. & Hansen, C. (2001). Application du cycle gestion qualité SMART des tests pédagogiques au cours d'Obstétrique et de Pathologie de la Reproduction des ruminants, équidés et porcins, Communication au 18ème Congrès de l'Association Internationale de Pédagogie Universitaire (AIPU), Les stratégies de réussite dans l'enseignement supérieur, Dakar, 5-7 avril 2001.

Chicago Board of Education (1999). *Standardized Test Preparation – Preparating Your High School Students to take Standardized Tests*, Chicago : Chicago Public Schools.

De Bal, R., De Landsheere, G., & Beckers, J. (1977). *Construire des échelles d'évaluation descriptives*. Bruxelles : Ministère de l'Education, Organisation des Etudes.

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.

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

Debry, M., Deltour, J.-J., Demeuse, M., Tanguy, E., Gilles, J.-L., Leclercq, D., Malherbe, E., Perée, F., & Poncin, P. (1999). *Observations sur les Etudes Universitaires à la Faculté (ŒUF) – Rapport de synthèse de la Commission Premier Cycle*. Liège : Université de Liège, Faculté de Psychologie et des Sciences de l'Education.

Didier, P., Faivre, M., Fauroux, R., Grandpierre, A., Millord, D. & Rousselet, M. (1966). *Le bouton du mandarin. L'école face à notre avenir*. Casterman, Centre d'Etudes Pédagogiques.

D'Ivernois J.-F., Gagnayre R. (1995). *Apprendre à éduquer le patient*. Paris : Vigot.

Dupont, P. & Ossandon, M. (1994). *La pédagogie universitaire*. Paris : Presses Universitaires de France.

Ebel, R.L. (1965). Confidence–Weighting and Test Reliability, *Journal of Educational Measurement*, 2, 49-57 B.

Ebel, R.L. (1979). Using tests to improve learning, *Arithmetic-Teacher*, v27, n3, p10-12.

Georges, F., Gilles, J.-L., Pirson, M., Simon, F. & Leclercq, D. (2001). Les feedbacks aux étudiants et aux

enseignants du projet MOHICAN, in Leclercq, D. (Ed.), *Le premier des MOHICANs – Une recherche-action de Monitoring Historique des CANdidatures*, Rapport relatif au contrat 798363 entre le Conseil Inter Universitaire Francophone (CIUF) et l'Université de Liège. Bruxelles : CIUF (à paraître).

Gilles, J.-L. (1995). Entraînement à l'autoévaluation : une comparaison filles/garçons à l'université, *Actes du Colloque de l'AIPU « Enseignement supérieur : stratégies d'enseignement appropriées »*, Hull : Université du Québec à Hull, pp. 159-166.

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., Gilles, J.-L., Jans, V. & Leclercq, D. (Eds), *Actes du 15ème Colloque de l'Association Internationale de Pédagogie Universitaire (A.I.P.U.)*. Liège : Affaires Académiques de l'Université de Liège, pp. 311-326.

Gilles, J.-L. (1998a). Apports des mesures métacognitives lors d'un test de compréhension d'un article scientifique, in Depover & Noël (Eds), *Approches plurielles de l'évaluation des processus cognitifs*. Mons : Université de Mons-Hainaut, pp. 19-30.

Gilles, J.-L. (1998b). Mise en œuvre de tests formatifs à l'aide de l'Internet, in Depover, C. & Noël, B.(Eds), *Approches plurielles de l'évaluation des processus cognitifs*. Mons : Université de Mons-Hainaut, pp. 193-204.

Gilles, J.-L. & Leclercq, D. (1995). Procédures d'évaluation adaptées à des grands groupes d'étudiants universitaires - Enjeux et solutions pratiquées à la FAPSE-ULG, in *Actes du Symposium International sur la Rénovation Didactique en Biologie*. Tunis : Université de Tunis.

Gilles, J.-L. & Melon, S. (2000). Comparaison de trois modalités de « testing » des compétences en français chez les étudiants médecins lors de leur première candidature à l'ULg, in Defays et al. (Eds), *La maîtrise du français du niveau secondaire au niveau supérieur*. Bruxelles : Editions De Boeck, pp. 161-178.

Gilles, J.-L., Bourguignon, J.-P. et Detroz, P., (2000). Les questionnaires à choix multiple : utilisation pour l'enseignement en groupe avec boîtiers électroniques. *Revue Médicale de Liège*, 55 : 12, pp. 1047-1050.

Gilles, J.-L., Collet, M., Debry, M., Denis, B., Etienne, A.-M., Geuzaine, C., Jans, V., Leclercq, D., Lejeune, M. & Paheau, C. (1998). *Evaluation des enseignements en 1ère et 2ème candidatures, année académique 1997-1998 - Rapport de synthèse pour le Conseil de Faculté du 10 novembre 1998*. Liège : Université de Liège, Faculté de Psychologie et des Sciences de l'Education.

Gilles, J.-L., Poncin, P., Ruwet, J.-C. et Leclercq, D. (1999). Les travaux dirigés virtuels d'Anthropologie biologique – Bilan d'une première utilisation, in J.-P. Béchard et D. Gregoire (Eds), *Apprendre et enseigner autrement*, Montréal : Ecole des Hautes Etudes Commerciales, Vol. 1, pp. 294-307.

Hunt, D. (1977). *The human self assessment process. Study II : The effects of the number of self-assessment categories on acquisition*. Interim Report from U.S. Army Research Institute for the Behavioral and Social Sciences Grant #DAHC19-76-G-002, New Mexico State University, Las Cruces, NM.

Jans & Leclercq (1999). Mesurer l'effet de l'apprentissage à l'aide de l'analyse spectrale des performances, in C. Depover & B. Noel (Ed.), *Evaluation des compétences et des processus cognitifs.* Bruxelles : De Boeck, pp. 303-317.

Karraker R.J. (1967). Knowledge of results and incorrect recall of plausible multiple choice alternatives, Journal of Educational Psychology, 58, 11-14.

Laveault, D. & Gregoire, J. (1997). *Introduction aux théories des tests en sciences humaines*. Bruxelles : De Boeck.

Laveault, D. (1999). *Autoévaluation et régulation des apprentissages*. in C. Depover & B. Noel (Ed.), *Evaluation des compétences et des processus cognitifs.* Bruxelles : De Boeck, pp. 57-79.

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

Leclercq, D. (1982). Confidence marking, its use in testing, in B. Choppin & N. Postlethwaite (eds.), *Evaluation in Education : International Review Series*, Oxford : Pergamon, vol. 6, n°2, pp. 161-287.

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

Leclercq, D. (1987). *Qualité des questions et signification des scores avec application aux QCM*. Bruxelles : Labor.

Leclercq, D. (1993). Validity, Reliability and Acuity of Self-Assessment in Educational Testing, in Leclercq, D. et Bruno, J. (Eds), *Item Banking : Interactive Testing and Self-Assessment*, NATO ASI Series. Heidelberg : Springer Verlag, pp. 113-131.

Leclercq, D. & Gilles, J.-L. (1993). Hypermedia : Teaching Through Assessment. In D. Leclercq et J. Bruno (Eds), NATO ASI Series, *Item Banking: Interactive Testing and Self Assessment*. Berlin: Springer Verlag, Vol. 112, pp. 31-47.

Leclercq, D. & Gilles, J.-L. (1994). GUESS, un logiciel pour entraîner à l'auto-estimation de sa compétence cognitive, in A. Dumont et J. Weber (Eds), *Actes du 3ème colloque international ESIEE* - Marne-La-Vallée *« QCM et questionnaires fermés »*. Paris : Université Paris 7 – Denis Diderot, Laboratoire d'ingénierie didactique, pp.137-158.

Leclercq, D. & Gilles J.-L. (1995). Le kaléidoscope des techniques de questionnement, *Colloque National de l'Association Internationale de Pédagogie Universitaire (A.I.P.U.),* Colonster-Liège, *22 septembre 1995*.

Leclercq, D. & Gilles, J.-L. (2001). Techniques de mesure dans l'autoévaluation (dix tecniques d'auto-estimation de la qualité de ses réponses), in G. Figari & M. Achouche (Eds), *L'activité évaluative réinterrogée – Regards scolaires et socioprofessionnels*. Bruxelles : Editions De Boeck, pp. 134-142.

Miller, G.-A. (1956). The magical number of seven, plus minus two, *Psychological Review*, Vol 63, 81-97.

Nightingale, P. & O'Neil, M. (1994). *Achieving quality learning in higher education*. London : Kogan Page.

Nitko, A. (1996). *Educational Assessment of Students*. Englewood Cliffs : Merill, second edition.

Passeron, J.C. (1970). Sociologie des examens. *Education et Gestion*, 2, 6-16.

Piaget, J. (1969). *Psychologie et pédagogie*. Paris : Denoël.

Pieron, H., Reuchlin, M. & Bacher, F. (1962). Une recherche expérimentale de docimologie sur des examens oraux de physique au niveau du baccalauréat de mathématiques, *Biotypologie*, 23, 48-73.

Pieron, H. (1963). *Examens et docimologie*. Paris, Presses Universitaires de France.

Shufford, A. (1993). In Pursuit of the Fallacy : Resurrecting the Penalty, in D. Leclercq & J. Bruno (Eds), *Item Banking : Interactive Testing and Self-Assessment*, NATO ASI Series, Berlin : Springer Verlag, pp. 76-98.

Shufford, A., Albert, A. & Massengil, N.E. (1966). Admissable probability measurement procedures, *Psychometrika*, 31, 125-145.

Van Naerssen, R.F. (1962). A scale for measurement of subjective probability, *Acta Psychologica*, 20, 2, 159-166.