PROCEDURES D'EVALUATION ADAPTEES A DE GRANDS GROUPES D'ETUDIANTS UNIVERSITAIRES : ENJEUX ET SOLUTIONS PRATIQUEES A LA FAPSE-ULG[1]

Jean-Luc GILLES[2] et Dieudonné LECLERCQ[3]
Faculté de Psychologie et des Sciences de l'Education - Université de Liège - Belgique
Tél. : + 32 - 41 - 66.20.78 - Fax : + 32 - 41 - 66.29.53 - Internet : jlgilles@vm1.ulg.ac.be

Conférence présentée par Jean-Luc GILLES
au Symposium International sur La Rénovation Didactique en Biologie
Tunis, novembre 1995

[1] Faculté de Psychologie et des Sciences de l'Education de l'Université de Liège en Belgique.
[2] Jean-Luc GILLES est assistant facultaire chargé des évaluations ayant recours aux moyens modernes tels que la lecture optique de marque et la passation de test interactifs sur ordinateurs à la FAPSE-ULG.
[3] Le professeur Dieudonné LECLERCQ dirige le Service de Technologie de l'Education de la FAPSE-ULG.

1. QUATRE PROBLEMES RECURRENTS EN EVALUATION

1.1 L'évaluation systématique de tous les processus mentaux ou comment augmenter la validité des épreuves

Selon DE LANDSHEERE (1980), valider une épreuve consiste à «*apporter la preuve (...) que l'examen fournit une évaluation correcte de ce qu'il prétend mesurer ou prédire* ». En d'autres termes, il s'agit de se poser la question : *les scores des étudiants représentent-ils bien ce que les enseignants veulent mesurer ?*

Répondre à cette question implique que la clarté ait été préalablement faite sur ce qui doit être mesuré, que les questions de l'examen soient clairement mises en lien avec les objectifs du cours.

La plupart du temps, les enseignants évaluent des processus mentaux appliqués à des contenus, or, s'il est aisé de mettre en évidence les contenus enseignés, il n'en va pas toujours de même en ce qui concerne les processus mentaux. Il existe cependant des outils qui permettent de clarifier ces processus mentaux. Ainsi la taxonomie de BLOOM (1969) souvent utilisée pour classifier différents niveaux de performances, est utile pour distinguer : (1) la connaissance ou restitution de mémoire, (2) la compréhension ou interprétation correcte de données, de concepts et de raisonnements, (3) l'application de principes à la solution de cas classiques, (4) l'analyse ou détection de problèmes et les classifications, (5) la synthèse, c'est à dire expression et (re)formulations, (6) l'évaluation ou le jugement sur base de critères personnels.

Bon nombre d'enseignants souhaitent évaluer la compréhension ou le jugement critique à propos de la matière qu'ils enseignent. Cependant, dans les évaluations habituellement pratiquées, seuls les niveaux (1) connaissance, (3) application et (5) synthèse sont mesurés systématiquement, c'est à dire de façon consciente, organisée et de telle manière que tous les étudiants se voient poser un échantillonnage de questions relevant de ces niveaux.

Comment mesurer aussi systématiquement les trois autres niveaux de la taxonomie de BLOOM que sont (2) la compréhension, (4) l'analyse et (6) l'évaluation ? Nous verrons plus loin comment ce défi a été affronté à la FAPSE-ULG à l'aide de plusieurs procédures combinées :

- les Questions à Choix Multiple (QCM) avec Solutions Générales Implicites (SGI) qui autorisent, en plus des solutions habituellement proposées, les quatre possibilités suivantes : Rejet (aucune solution proposée n'est correcte), Toutes (toutes sont correctes), Manque (il manque des données dans *l'énoncé pour que l'on puisse choisir UNE solution comme correcte)*, Absurdité (il y a une contre-vérité dans l'énoncé *à dénoncer en priorité !);*
- les évaluations à livre ouvert;
- les questions en deux volets (double check);
- les degrés de certitude.

1.2 La garantie d'une correction objective ou comment augmenter la fidélité des mesures

Dans quelle mesure un correcteur peut-il prétendre qu'un travail corrigé et classé dans la catégorie « excellent » bénéficierait de la même mention s'il était à corrigé dans d'autres conditions (autres correcteurs ou un mois plus tard) ? Les problématiques liées à la subjectivité de la correction ont été l'objet d'un courant de docimologie dite « négative » ou « critique » mené par PIERON (1963). Ce dernier et d'autres chercheurs à sa suite ont relevé une série de biais d'évaluation tels que l'inconstance d'un même évaluateur[4] et la discordance

[4] Par exemple *l'effet d'ancrage* observé par BONNIOL (1972) lors de la correction de travaux de valeur moyenne parmi lesquels il introduit des ancres (copies de valeur soit excellente, soit médiocre) et qui provoquent un effet de contraste sur les travaux suivants.

entre évaluateurs. Cette dernière est illustrée, entre autres, par AGAZZI (1967) qui observe à l'occasion de la correction des copies d'un baccalauréat par 6 correcteurs, 70 % des compositions françaises qui sont tantôt admises par les uns et tantôt refusées par les autres. PIERON & Al. (1962) estiment qu'il faudrait 16 correcteurs pour stabiliser les notes en physique, 78 correcteurs en composition française et 127 en dissertation philosophique...

Des procédures d'évaluation automatisées visent à garantir la fidélité des mesures. Les QCM permettent d'échapper à la subjectivité des correcteurs et contribuent ainsi à augmenter la fidélité des évaluations; de plus, la simplicité de correction autorise un traitement informatisé rapide grâce à la lecture optique de marques[5] ou au questionnement interactif (les questions sont alors posées via l'écran d'un ordinateur ce qui permet notamment un feed-back après chaque réponse fournie ou/et immédiatement en fin de test). Une autre source de fluctuations est liée à l'impossibilité habituelle d'exprimer son doute ou la sûreté de ses connaissances.

1.3 La sensibilité diagnostique des techniques de questionnement

Comment diagnostiquer avec précision les difficultés d'apprentissage, les processus maîtrisés et ceux qui ne le sont pas ? Habituellement pour développer des procédures de diagnostic et de remédiation l'évaluateur recourt à des échelles d'évaluation descriptives. Par exemple, il pose une question ouverte et il la corrige en fonction de *x* critères opérationnalisés. La tâche qui est ainsi demandée à l'évaluateur est très complexe, vu l'effort d'analyse, nécessaire pour séparer les différentes catégories de réponses. En outre, il n'est pas possible de procéder à un diagnostic univoque à l'aide d'UNE SEULE questions. Il est bien connu, en effet, qu'on ne peut dire d'une question qu'elle mesure A COUP SUR la connaissance, ou la compréhension, ou l'analyse car CELA DEPEND de ce que l'étudiant maîtrise par ailleurs. Le « recoupement » de plusieurs réponses à plusieurs questions permet lui, un diagnostic, comme cela se passe en médecine : la température à elle seule ne permet pas d'identifier la maladie, mais COMBINEE avec d'autres observations, elle le permet.

C'est le principe qui a présidé à la conception de « double check » (LECLERCQ, 1993), une procédure d'évaluation interactive, qui consiste à poser une question en deux volets *prim* et *bis*. L'étudiant reçoit une première (*prim*) question (QCM-SGI) où la réponse correcte attendue peut, par exemple, être « 8. Manque de données dans l'énoncé ». Après avoir répondu, l'étudiant reçoit la réponse puis la deuxième partie de la question (*bis*), par exemple : « quelle donnée manque ?». Suivent à nouveau une série de propositions. Les performances des étudiants se présentent alors selon différents cas de figure qui peuvent ensuite donner lieu à des procédures de remédiation adaptées selon le diagnostic. Dans le cadre de notre exemple :

	Volet *prim* (analyse)	**Volet *bis* (compréhension)**	**Diagnostic :**
Compétence totale	Réussite	Réussite	Analyse et compréhension correctes du problème.
Compétence partielle	Echec	Réussite	Manque de vigilance (mais compréhension).
Compétence partielle	Réussite	Echec	Incompréhension du problème. La solution choisie dans le volet bis peut indiquer l'erreur de raisonnement.
Incompétence	Echec	Echec	Les solutions choisies dans la partie *prim* et *bis* peuvent donner des indications quant aux causes de l'échec.

1.4 La praticabilité des dispositifs d'évaluation ou « la survie des correcteurs... »

[5] Les étudiants répondent en cochant leurs réponses sur des feuilles spéciales qui sont ensuite lues par un dispositif de lecture optique qui peut traiter jusqu'à 6400 copies à l'heure.

Garantir l'objectivité des corrections, l'évaluation systématique des processus mentaux en jeux ainsi que l'augmentation de la sensibilité diagnostique impliquent dans le cadre d'examens classiques[6] que l'on y consacre de l'énergie et du temps. Il est possible de poser une question ouverte à réponse écrite longue notée selon *x* points de vue différents à l'aide de *y* critères préalablement opérationnalisés pour chacun, et il en résulte une correction complexe. Or, comme dans la plupart des universités européennes (GIBBS, JENKINS & al., 1992), on assiste aujourd'hui à une explosion du nombre d'inscriptions d'étudiants à la FAPSE-ULG[7], et, si on multiplie par le nombre d'étudiants le temps passé à corriger un examen dans ces conditions on en arrive rapidement à un constat d'impraticabilité.

Depuis 1994, la FAPSE-ULG, sous l'impulsion de son Doyen, le Professeur DE KEYSER, a décidé d'aider les enseignants du premier cycle (là où les étudiants sont les plus nombreux) en mettant à leur disposition, dans le cadre du Centre d'Auto-Formation et d'Evaluation Individualisé Multimédia de la Faculté de Psychologie et des Sciences de l'Education[8] (CAFEIM-FAPSE), un dispositif ayant recours aux techniques d'évaluation informatisées, notamment par QCM-SGI avec degrés de certitude ... et un scientifique spécialisé (J.-L. GILLES). Le soutien apporté aux enseignants porte sur la rédaction des questions, la gestion de la banque de questions, la mise en page des différentes formes de questionnaires, la duplication, l'aide à la surveillance des épreuves, la correction à l'aide de la lecture optique, l'analyse statistique des résultats, le feed-back aux professeurs et aux étudiants.

De plus en plus d'enseignants de la FAPSE ainsi que d'autres facultés[9] de l'université de Liège font appel au dispositif de CAFEIM-FAPSE. Dans plusieurs examens, deux types de questions sont proposés en symbiose. D'une part, un grand nombre de QCM-SGI (en général une trentaine) avec degrés de certitudes offrent une série d'avantages décrits par LECLERCQ (1986) : représentativité de l'échantillon des questions, simplicité de correction, objectivité de la correction, possibilité d'évaluer systématiquement les niveaux de compréhension, d'analyse et d'évaluation/jugement. D'autre part, un petit nombre de questions ouvertes (en général une ou deux) permettent d'évaluer l'esprit de synthèse, l'originalité, la créativité, la capacité à organiser une réponse, ce qui n'est pas possible en ayant recours uniquement aux QCM fussent-elle SGI.

2. SOLUTIONS TECHNOLOGIQUES MISES EN OEUVRE A LA FAPSE-ULG

2.1 Les modalités de questionnement privilégiées dans le cadre de notre système d'évaluation informatisé

LECLERCQ et GILLES (1995b) distinguent sept facettes d'évaluation dans le cadre d'un modèle intitulé *Kaléidoscope du questionnement* : les objectifs (O), les procédures (P) (par exemple, la disponibilité d'un feed-back juste après la réponse), les différents types de questions (Q), les modalités de réponses (R), les séquences (S) (par exemple, la liaison de plusieurs questions à l'aide du *double check*), le « training » (T) c'est à dire l'entraînement à la situation d'examen, et, les unités de tarification (U) ou points accordés. Combinées entre elles, les composantes de ces différentes facettes produisent un grand nombre de situations d'examen. Notre système informatisé en privilégie les composantes suivantes :

- (0) Les objectifs liés aux processus mentaux d'évaluation/jugement, d'analyse/détection et de compréhension/interprétation;

[6] Nous entendons par *examens classiques*, les évaluations ayant recours aux questions ouvertes à réponse construite longue posées soit à l'oral, soit à l'écrit.

[7] D'environ 200 en 1986-1987, le nombre d'inscriptions en première candidature de la FAPSE-ULG est passé à 400 en 1994-1995.

[8] CAFEIM-FAPSE a été créé en 1988 par le Service de Technologie de l'Education (Prof. D. LECLERCQ).

[9] Notamment les facultés des Sciences Appliquées, des Sciences Vétérinaires, de Droit et l'Institut Supérieur des Langues Vivantes.

- (P) Dans le cadre des évaluations interactives où les questions sont posées via l'écran d'un ordinateur, l'étudiant peut recevoir la communication de la solution correcte et, s'il le désire, de son score immédiatement après sa réponse. La modification des réponses (double passage) est aussi possible;
- (Q) Les Questions à Choix Multiple avec Solutions Générales Implicites (QCM-SGI);
- (R) Les réponses sont obligatoirement accompagnées d'un degré de certitude. Rappel de l'échelle : 0 = de 0% à 25%, 1 = de 25% à 50%, 2 = de 50% à 75%, 3 = de 75% à 85%, 4 = de 85% à 95% et 5 = de 95% à 100%. Cette échelle est fondée sur les travaux de D. LECLERCQ (1983, 1993) concernant les limites de la précision humaine en matière d'estimation du doute et de la certitude.
- (S) Dans le cadre des évaluations interactives, les questions sont liées entre elles (double check) et l'ordre de présentation des « composés » de questions, est au choix de l'étudiant;
- (T) Deux types d'entraînement sont offerts. Les « quizzes » (voir ci-après) constituent un entraînement fin de cours collectif aux QCM-SGI avec degrés de certitude. Le logiciel GUESS (voir ci-après) permet un entraînement individuel à la gestion des degrés de certitude;
- (U) Le barème des tarifs lié aux degrés de certitude est calculé conformément à la théorie des décisions.

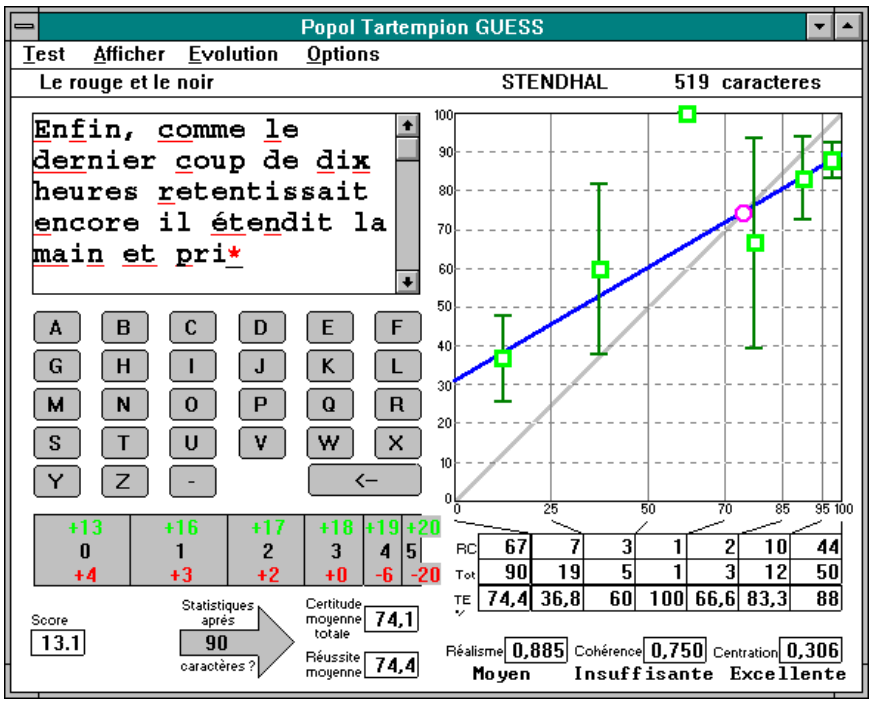
2.2 L'entraînement à la métacognition avec GUESS[10]

Dans leur grande majorité, les étudiants du premier cycle de la FAPSE-ULG n'ont jamais rencontré au cours de leur cursus scolaire de procédure d'évaluation ayant recours aux examens à livre ouvert avec degrés de certitude, solutions générales implicites et double check informatisé. Dès lors, il est indispensable de les y entraîner avant le premier examen.

L'entraînement à l'auto-estimation de sa compétence cognitive à l'aide du logiciel GUESS a été décrit en détail par LECLERCQ & GILLES (1994). Chaque étudiant s'entraîne individuellement à l'utilisation des degrés de certitude à l'aide d'un jeu où il doit deviner les lettres successives d'un texte d'au moins cent lettres (inspiré de SHANNON, 1951 & ATTNEAVE, 1959). Le joueur effectue une prédiction en tapant une lettre qu'il accompagne de la probabilité subjective de réussite exprimée à l'aide d'un degré de certitude. Il est ensuite informé de la réponse correcte qui s'affiche dans la zone réservée au texte. Lettre par lettre, le texte s'affiche ainsi à l'écran. Evidemment, le début des mots est plus difficile à deviner que leur fin (c'est ce que SHANNON voulait démontrer). Après un nombre donné[11] de réponses, un graphique de réalisme se construit dans le coin supérieur droit de l'écran. L'étudiant peut y observer, pour chacun des degrés de certitude utilisés (en abscisse) le taux de réussite (en ordonnée) atteint. Idéalement, ces taux d'exactitude devraient se placer sur la diagonale (les valeurs attendues) et non plus haut (sous-estimation) ou plus bas (surestimation). Les taux (ou pourcentages) de réussite sont encadrés de leur erreur de mesure. Voici l'écran tel que le voit le joueur :

[10] GUESS a été conçu par le Prof. LECLERCQ et programmé dans le langage ToolBook par M. HURARD, licencié en informatique. Il est possible d'obtenir des licences d'utilisation du logiciel GUESS à des conditions « éducation » en s'adressant au Service de Technologie de l'Education, tél. +32-41-66.20.72, fax +32-41-66.29.53.

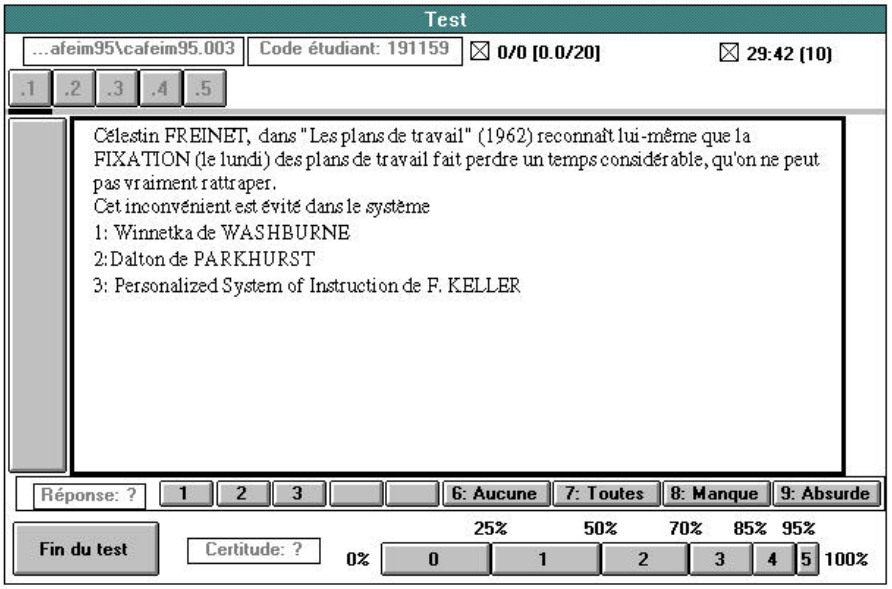
[11] Ce nombre est paramétrable à volonté (ex : au bout de 10 questions, ou 12, ou 20, etc.)



Le petit cercle constitue le «centre de gravité » du graphique de réalisme : ses coordonnées sont d'une part la certitude moyenne (ici 74,1%) et d'autre part l'exactitude moyenne (ici 74,4%). Dans le cas qui nous occupe, cette centration moyenne quasi parfaite résulte en effet de compensations de surestimations par des sous-estimations.

2.3 L'examen informatisé interactif avec WINCHECK[12]

WINCHECK est un logiciel d'évaluation interactive fonctionnant dans l'environnement Windows et qui permet l'utilisation du «double check ». L'évaluateur peut se constituer des banques de questions (QCM-SGI) et créer des tests en sélectionnant les questions en fonction de critères liés aux contenus, aux processus mentaux et aux objectifs. Voici l'écran d'une question posée à l'aide de WINCHECK :



[12] Le logiciel WINCHECK a été développé au Service de Technologie de l'Education de l'Université de Liège dans le cadre du programme européen CERT-EUROFORM.

Avant le premier examen ayant recours à ce logiciel, tous les étudiants du premier cycle sont invités à s'entraîner (exercice sans sanction) à l'utilisation du programme WINCHECK et de la procédure «A livre ouverts ». Le test d'entraînement est composé d'une dizaine de questions portant sur un magazine.

Les étudiants prennent rendez-vous pour l'examen en réservant un ordinateur à CAFEIM-FAPSE dans une plage horaire qui leur convient.

2.4 L'évaluation en grand auditoire : lecture optique et logiciel CERT

Une façon plus habituelle d'évaluer les étudiants consiste à les rassembler tous dans un auditoire. Ils reçoivent alors les QCM-SGI de l'épreuve sur des questionnaires accompagnés de feuilles de réponses spéciales où ils cochent les cases correspondant à leurs solutions choisies et à leurs degrés de certitude. Pour certains cours, les évaluateurs autorisent de brèves justifications des réponses sur des feuilles ad hoc. Il est en outre convenu que le(s) correcteur(s) ne lira(ont) que les commentaires concernant les réponses incorrectes. La justification ne peut donc QUE bénéficier à l'étudiant.

Une fois lues les feuilles destinées à la lecture optique, les réponses sont traitées par le logiciel CERT, réalisé en 1991 sous les auspices de la Commission des Communautés européennes dans le cadre du programme EUROTECNET (BOXUS et al. , 1991).

Le logiciel CERT fournit l'analyse de chaque question de l'épreuve en trois lignes par question. Pour chaque solution proposée (de 1 à 9 pour les QCM-SGI) sont en effet fournis : (1) le pourcentage d'étudiants qui ont choisi la proposition, (2) le coefficient *r*.bis[13] et (3) la certitude moyenne. Voici comment ce présente l'écran ou le listing (les statistiques concernant la réponse correcte sont encadrées de deux barres verticales).

Certm

Processus : Pas de sélection. Matière : Pas de sélection.

	SOL: 0	1	2	3	4	5	6	7	8	9
(1)	Q 26	0.00	26.40	3.37	7.30	**33.71**	5.62	11.24	1.69	1.12	9.55
(2)	Rbis	0.00	0.01	0.11	-0.27	**0.25**	-0.03	-0.03	-0.04	-0.12	-0.14
(3)	Cmoy	0.00	41.44	32.50	42.31	**49.21**	33.50	37.63	29.17	45.00	49.71
	Q 27	0.56	17.42	1.69	**24.16**	8.99	10.67	18.54	14.04	2.81	1.12
	Rbis	-0.18	0.07	0.00	**0.09**	-0.05	-0.13	0.04	0.00	-0.09	0.05
	Cmoy	12.50	37.26	37.50	**38.95**	36.41	31.84	44.09	44.30	36.50	12.50
	Q 28	0.00	1.12	2.81	**44.94**	26.97	10.67	8.99	0.00	1.69	2.81
	Rbis	0.00	-0.08	-0.10	**0.25**	-0.14	-0.14	0.06	0.00	-0.03	-0.04
	Cmoy	0.00	25.00	12.50	**59.03**	53.13	40.00	46.41	0.00	46.67	32.00
	Q 29	0.00	3.93	2.25	3.93	8.99	**68.54**	3.93	0.00	0.56	7.87
	Rbis	0.00	-0.24	-0.09	-0.15	-0.11	**0.39**	-0.13	0.00	-0.09	-0.12
	Cmoy	0.00	42.50	56.25	40.36	47.97	**61.37**	51.79	0.00	60.00	56.96

Frappez une touche pour continuer

L'analyse d'une épreuve à l'aide de ce dispositif peut amener l'évaluateur à supprimer une question, par exemple lorsque la réponse correcte récolte un *r*.bis négatif élevé et les distracteurs des *r*.bis positifs. Il arrive que l'enseignant décide d'accepter le choix d'un distracteur comme réponse correcte au vu des *r*.bis et des commentaires de justification des étudiants. La question doit alors être remaniée avant d'être réintroduite dans un examen ultérieur. Il est également possible d'exporter les résultats des évaluations dans les logiciels EXCEL et STATISTICA en vue de traitements plus poussés.

2.5 L'entraînement par les QUIZZES en fin de cours

Les étudiants du premier cycle sont entraînés à cette situation d'examen dès le premier semestre à l'aide de la méthode des quizzes, c'est à dire quelques questions (entre 5 et 8) posées en fin de cours, corrigées pour le cours suivant avec feed-back personnalisé distribué à chaque étudiant. Les quizzes permettent également au professeur de commenter (lors du cours suivant) les statistiques de résultats de l'ensemble des étudiants, les erreurs les plus fréquentes et éventuellement de réexpliquer une matière mal comprise.

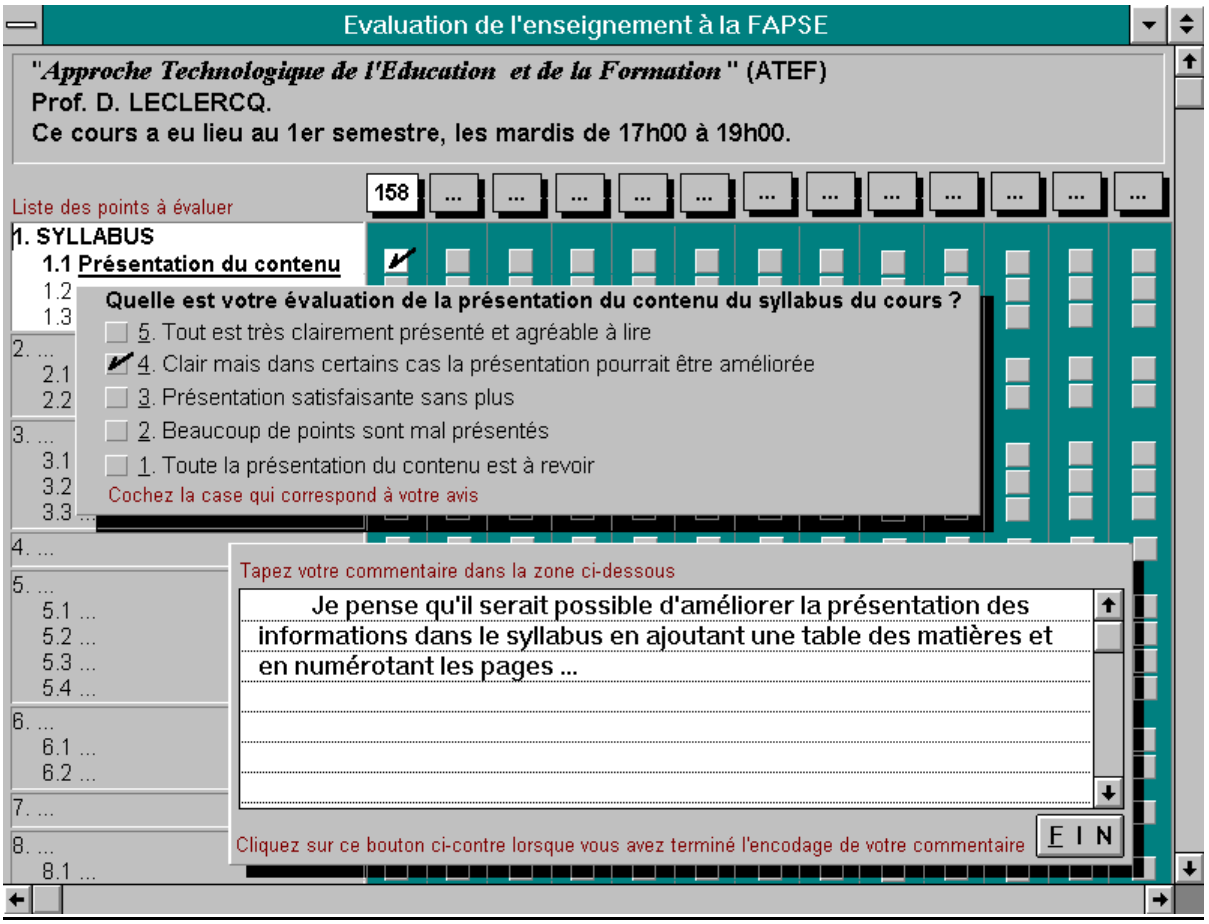
2.6 Le dialogue pédagogique systématisé et l'évaluation des enseignements

Des commissions mixtes (étudiants/professeurs) de la FAPSE ont étudié les modalités de recueil des avis des étudiants sur l'enseignement qu'ils reçoivent et de renvoi des feed-back aux professeurs. Dans ce but, le logiciel DPS (Dialogue Pédagogique Systématisé) EVALENS a été conçu par LECLERCQ et GILLES (1995a) et programmé par HURARD.

L'originalité du programme réside dans la possibilité qui est offerte aux étudiants d'accompagner chaque réponse exprimée sur une échelle d'évaluation par un commentaire. Une fois récoltés et triés, ces

[13] Le coéfficient *r*.bis est une corrélation entre le choix des étudiants à chaque QCM et le score total à l'épreuve obtenu.

commentaires permettent à l'enseignant d'avoir une idée précise sur la nature des critiques positives ou négatives qui lui sont faites.

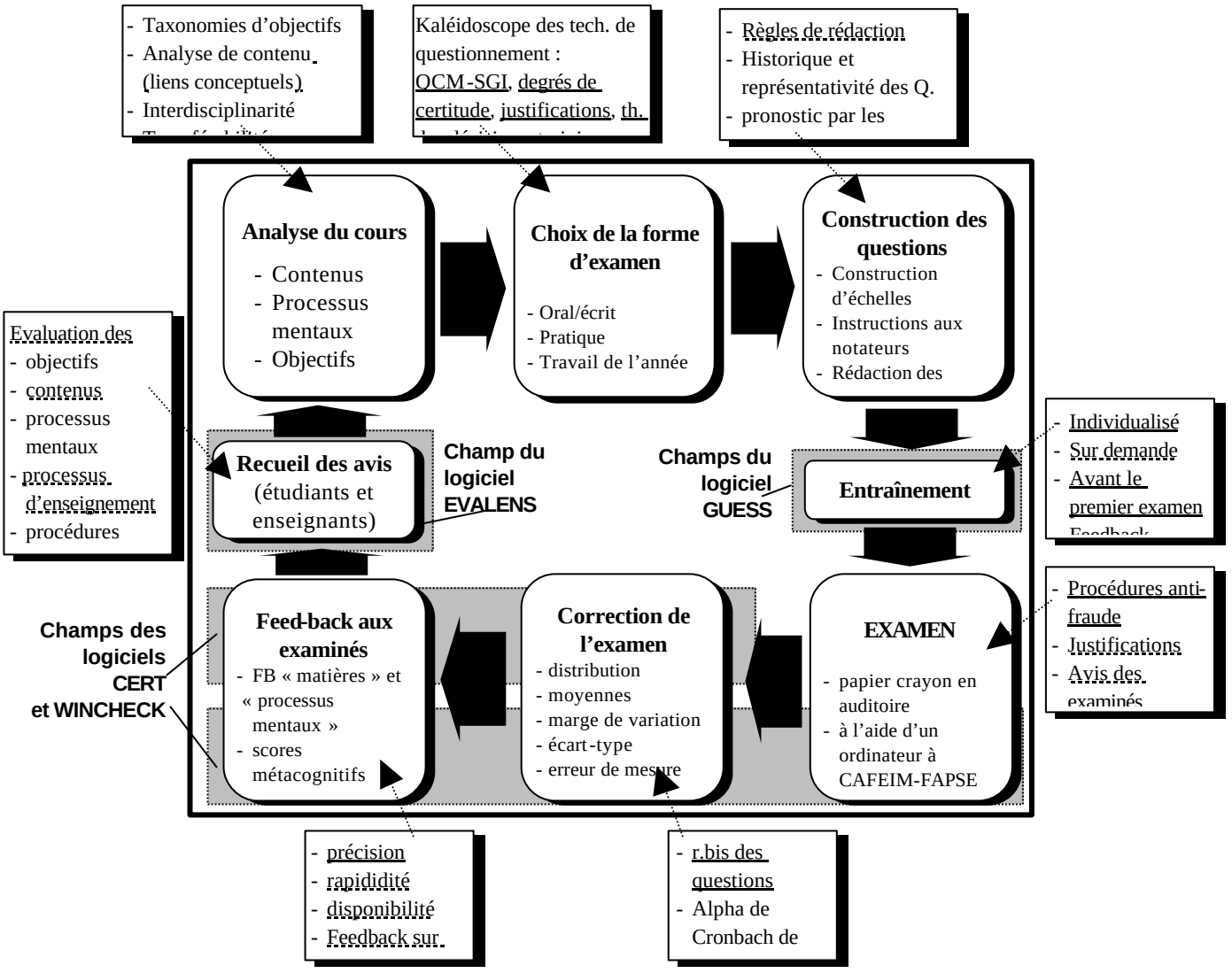


Le logiciel ne vise pas la comparaison entre professeurs, il ne comporte pas de fonction ayant pour but de comparer publiquement les enseignements entre-eux. En plus des commentaires triés, chaque professeur recevra pour chaque rubrique la moyenne générale ainsi que sa moyenne personnelle et l'écart-type, mais restera dans l'ignorance des « scores » de ses collègues.

3. VERS UN MODELE GLOBAL DE GESTION DES EXAMENS

3.1 Modéliser le processus de construction des examens

Inspiré du schéma des grandes phases de construction d'un examen conçu par DE LANDSHEERE (1980), le modèle présenté ci-après met l'accent sur les particularités de la procédure d'évaluation mise en place dans le cadre de CAFEIM-FAPSE. Une série de recommandations en vue d'augmenter la validité et la fidélité des examens sont présentées dans les cadres rectangulaires ombrés qui entourent le schéma. Les recommandations soulignées en continu sont mise en place de façon systématique dans nos évaluations. Celles qui sont soulignées en pointillés sont, à notre avis, encore trop peu suivies par les enseignants, certaines ne le sont pas du tout et ne sont pas soulignées dans les cadres. Les plages hachurées délimitent les champs d'action des logiciels décrits succinctement ci-avant, et ceux en creux, désignent les champs qui restent à approfondir.



3.2 La rédaction des questions (assurer la qualité *a priori*)

En ce qui concerne la construction des questions à choix multiple, nous recommandons aux enseignants l'utilisation d'une série de principes définis par LECLERCQ (1986). Souvent, nous effectuons à la demande des professeurs une relecture « formelle » de leurs QCM suivie d'un entretien où nous discutons des pistes à suivre en vue d'améliorer la qualité des questions. Cette façon de procéder peut être combinée avec

celle des enseignants de la Faculté de Médecine de la Rijksuniversiteit Limburg de Maastricht[14] qui proposent à leurs étudiants des tests de 300 QCM tous les trois mois sur toute la matière de médecine (le niveau d'exigence requis dépend du niveau d'étude de l'étudiant). Ces enseignants vérifient la pertinence de leurs questions en demandant à leurs collègues d'y répondre et de les commenter. Les questions utilisées dans ce type de système doivent être gérée dans le cadre d'une ou plusieurs banques où chaque QCM est accompagnée de son « histoire » : les examens où elle fut posée, les scores et statistiques qu'elle récolta, les améliorations successives, etc.

3.3 L'analyse des qualités psychométriques des questions (vérifier la qualité *a posteriori*)

Les examinés ont des choses à dire sur les évaluations qu'ils subissent, il nous semble important de récolter leurs avis afin d'améliorer la validité et la fidélité des examens. La procédure des feuilles de justification permet à l'étudiant d'exprimer de façon brève (un cadre de réponse limite chaque commentaire à une cinquantaine de mots) son trouble face à certaines questions. Ces justifications subissent ensuite un tri : seuls sont lus les commentaires concernant les réponses incorrectes (afin d'alléger le travail de correction), ceci, après l'analyse des coefficients *r*.bis, amène souvent le professeur à changer les critères de correction d'une évaluation, par exemple, un distracteur s'avère correct et on accorde le point soit à tous les étudiants qui l'on désigné comme étant exact, soit seulement à ceux qui ont accompagné cette réponse d'un commentaire justifiant leur choix. Il arrive également qu'une question mal formulée et décelée par cette méthode soit éliminée, on améliore ainsi la fidélité de l'épreuve.

3.4 Les feed-back aux étudiants

Notre dispositif permet de donner aux étudiants un feed-back détaillé sur leurs performances à l'examen, des informations métacognitives sont aussi fournies, notamment un graphique de réalisme identique à celui auquel ils ont été confronté dans le cadre de l'entraînement GUESS (voir 2.2 ci-avant). Des progrès sont encore à réaliser, notamment au niveau de la rapidité et de la disponibilité du feed-back. Chaque étudiant de la FAPSE-ULG dispose d'une adresse d'accès au réseau internet et nous pensons que l'on pourrait utiliser ce réseau pour la diffusion des résultats après chaque évaluation. Internet ouvre également de nouvelles perspectives en ce qui concerne les évaluations à distance et à la demande.

3.5 Perspectives de recherches et développements

De nouveaux outils sont à créer et à intégrer. Ils devraient aider les enseignants à clarifier, non seulement les contenus à évaluer, mais aussi les processus mentaux et par là leurs objectifs. Ils devraient également contribuer à faire le lien avec la forme d'examen la plus appropriée. Cela participerait, nous semble-t-il, à l'amélioration de la validité (Cfr. 1.1 ci-avant).

Les aspects liés à l'interdisciplinarité ne sont pas à négliger : beaucoup de contenus de cours différents se rejoignent et c'est heureux. Quelques enseignants de la FAPSE, les professeurs CRAHAY (Service de Pédagogie Expérimentale), LECLERCQ (Service de Technologie de l'Education) et THIRION (Service de Méthodologie de l'Enseignement) en sont conscient et mettent en place un cours intégré d'*Introduction aux sciences de l'éducation*, chaque enseignant intervenant pour un tiers du cours. De telles pratiques devraient amener plus de cohérence dans les cursus et aider les étudiants à mieux percevoir les relations qu'entretiennent les différentes disciplines relevant d'un même domaine. Cela implique également de repenser l'évaluation : comment évaluer systématiquement la capacité des étudiants à faire les liens entre les différentes approches d'un même phénomène ?

[14] qui pratiquent l'Approche Basée sur les Problèmes (ABP) décrit par VAN DER VLEUTEN et WIJNEN (1990)

Comme dans le cadre des feuilles de justification des réponses, mais cette fois pour améliorer la validité, on devrait plus fréquemment demander aux étudiants leur avis sur la représentativité des questions. Comment perçoivent-ils les liens entre les questions posées à l'examen et les objectifs de l'enseignement qu'ils ont suivi ? Quelles sont leurs commentaires à ce sujet ? Récoltés systématiquement à l'occasion de chaque examen et pris en compte par les enseignants, ces avis amélioreraient la validité des épreuves.

De façon plus globale, en vue d'améliorer les cours, il est intéressant de récolter les avis à propos du processus d'enseignement et d'établir les liens entre les résultats, les caractéristiques des cours, les objectifs, les modalités d'évaluation et la perception qu'ont les étudiants et les enseignants de l'ensemble.

Des différences métacognitives liées à l'auto-estimation de ses compétences sont observées lors des évaluations. La tendance à la surestimation souvent décrite dans la littérature semble pouvoir se nuancer lorsqu'on compare les tendances à *sur* et *sous*-estimer ses compétences chez les filles et les garçons non entraînés à la gestion des degrés de certitude à l'aide du logiciel GUESS (GILLES, 1995a et 1995b). Ce type d'hypothèse de recherche est également étudié par DIRKZWAGER & al. (1993, 1995) dans le cadre d'une recherche interculturelle réunissant des chercheurs américains, australiens, espagnols, mexicains et néerlandais.

Enfin, le problème de la transférabilité des compétences acquises au cours des études et certifiées lors des évaluations doit, à terme, également être pris en compte. En ne nous donnant pas les moyens d'estimer la qualité des transferts nous nous privons d'informations primordiales pour un enseignement cohérent et connecté sur la réalité professionnelle.

* *
*

4. BIBLIOGRAPHIE

AGAZZI, A. (1967).Les aspects pédagogiques des examens, Strasbourg, Conseil de l'Europe, p. 119.

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

BLOOM, B. & al. (1969). Taxonomie des objectifs pédagogiques. I. Domaine cognitif, traduit par M. Lavallée, Montréal, Education Nouvelle.

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, 1972.

BOXUS & Al. (1991). Principes communs pour évaluer les résultats cognitifs de la formation. Commissions des Communautés européennes, programme Eurotecnet.

DE LANDSHEERE (1980). Evaluation continue et examens - Précis de docimologie, Bruxelles, Ed. Labor, Eucation 2000.

DIRKZWAGER, A. (1993). A computer environment to develop valid and realistic predictions and self-assessment of knowledge with personal probabilities. NATO ASI Series, Item Banking: Interactive Testing and Self Assessment, Berlin: Springer Verlag, 1993, Vol. 112, pp. 146-166.

DIRKZWAGER, A. (1995). Testbet version 1.01 instructions manual. Bussum, Computers in Education, 1995.

GIBBS, G., JENKINS, A. & al (1992). Teaching large classes in higher education - How to maintain quality with reduced resources, London: Kogan Page, 1992.

GILLES, J.-L. (1995a). Gender comparison of metacognition : realism in self-estimation in cognitive competency with university students, à paraître.

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

LECLERCQ, D. (1983). Confidence marking, its usein testing. Postlethwaite, Choppin (eds.) Evaluation in Education, Oxford : Pergamon, 1982, vol. 6, 2, pp. 161-287.

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

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

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

LECLERCQ, D. et GILLES J.-L. (1995a). EVALENS, Propositions pour un logiciel d'évaluation des enseignements. Commission d'évaluation des enseignements de la Faculté de Psychologie et des Sciences de l'Education, Université de Liège.

LECLERCQ, D. et GILLES J.-L. (1995b). Le kaléidoscope des techniques de questionnement. Actes de la journée de l'Association Internationale de Pédagogie Universitaire, Liège, à paraître.

PIERON, H. & Al. (1962). Une recherche expérimentale de docimologie sur les examens oraux de physique au niveau du baccalauréat de mathématique, in *Biotypologie*, mars-juin 1962, p. 5189.

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

SHANNON, C.E. (1951). Prediction and entropy of printed english. Bell Syst. Techn. J. 30, pp. 50-64.

VAN DER VLEUTEN C. et WIJNEN, W (1990). Problem-Based learning : Perspective from the Maastricht experience, Amsterdam, Thesis.