

**Colloque de l'AIPU
Section Communauté Française de Belgique¹
Colonster - Liège
22 septembre 1995**

sur

L'évaluation par QCM

**Le kaléidoscope
des techniques de questionnement**

Dieudonné LECLERCQ² et Jean-Luc GILLES³

Université de Liège
Faculté de Psychologie et des Sciences de l'Education
Sart-Tilman - Bâtiment B32
Fax : 32-41-66.29.53

¹ Association Internationale de Pédagogie Universitaire - Section "Communauté Française de Belgique".

² Professeur à l'ULG, Directeur du Service de Technologie de l'Education, FAPSE – d.leclercq@ulg.ac.be

³ Assistant facultaire à la Faculté de Psychologie et des Sciences de l'Education (FAPSE), chargé d'aider les enseignants de candidatures dans les évaluations informatisées de grands groupes – jl.gilles@ulg.ac.be

A. VALIDITE D'ABORD

"Les scores (mesures) représentent-ils (elles) bien ce que l'on prétend mesurer, tout ce que l'on veut mesurer et rien d'autre ?" C'est la question de la validité des mesures en évaluation. Or les formateurs veulent mesurer des processus mentaux appliqués à des contenus. Un repère commode et bien connu relatif aux processus mentaux est la taxonomie des objectifs cognitifs de B. BLOOM *et al.* (1956) :

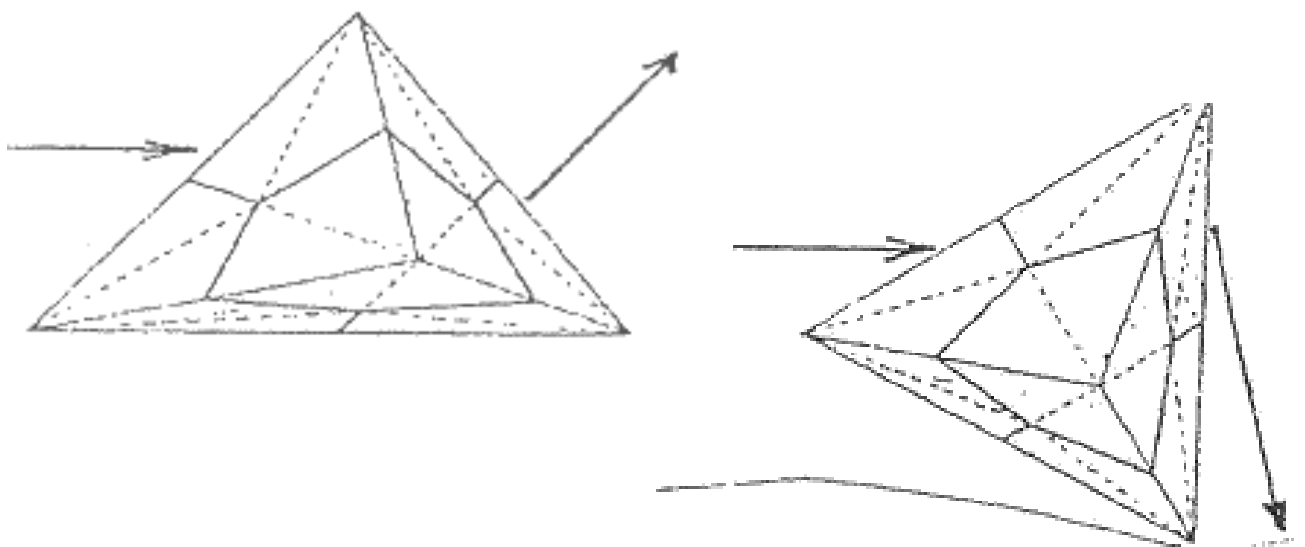
1. **connaissance** (de mémoire) mais dont il importerait de distinguer plus nettement la mémoire de rappel de la mémoire de reconnaissance;
2. **compréhension** (traduction);
3. **application** (des règles générales à des cas particuliers);
4. **analyse** (détection de pièces, d'anomalies);
5. **synthèse** (expression - formulation);
6. **évaluation** (jugement et prise de risque sous incertitude).

Force est de constater que dans l'enseignement (primaire, secondaire, supérieur), trois seulement de ces 6 niveaux sont systématiquement mesurés : les connaissances (récitations), l'application (problèmes classiques) et la synthèse (rédactions, ...).

Nous avons développé ailleurs (LECLERCQ *et al.*, 1993) les raisons d'évaluer tout aussi systématiquement, et avec les modalités de questionnement et d'interprétation appropriées, la compréhension, l'analyse, le jugement.

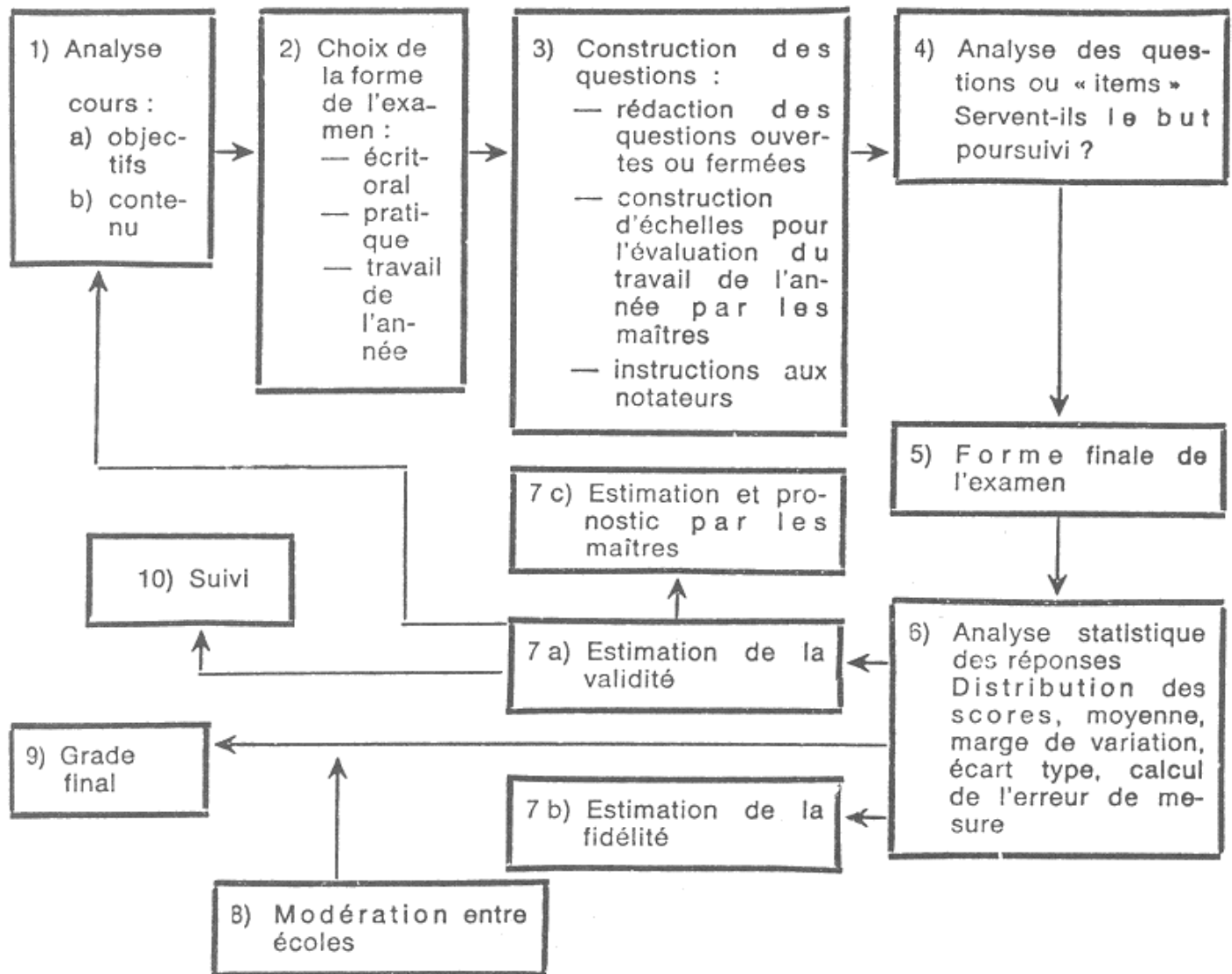
Les techniques d'interrogation, dont il sera largement question ci-après, sont au service de ces objectifs. Il importe de ne pas le perdre de vue.

Dans le "kaléidoscope" ci-après, ces objectifs sont placés dans la TOUTE PREMIERE FACETTE, et constituent en fait la source lumineuse (qui, passée à travers le "prisme de l'observation" donnera des mesures).



B. LA CONSTRUCTION D'UN EXAMEN

G. DE LANDSHEERE, dans son fameux ouvrage "Evaluation continue et examens. Précis de Docimologie" (1992, 6e éd., p. 65) nous donne une vue d'ensemble des opérations :



Le document qui suit ne traite que des opérations 2 et 3.

C. LES COMPETENCES SONT OBSERVEES A TRAVERS UN PRISME

Les performances (réciter, calculer, dessiner, raconter, etc.) servent de révélateurs des compétences, véritable enjeu de l'évaluation. Pour obtenir ces performances révélatrices, on peut attendre qu'elles se produisent spontanément mais, parce qu'on est pressé (à juste titre), le plus souvent on les sollicite, soit **implicitement** (l'évaluateur crée une situation à laquelle il s'attend que l'apprenant réagisse, mais sans le lui dire), soit **explicitement**, par des questions directes.

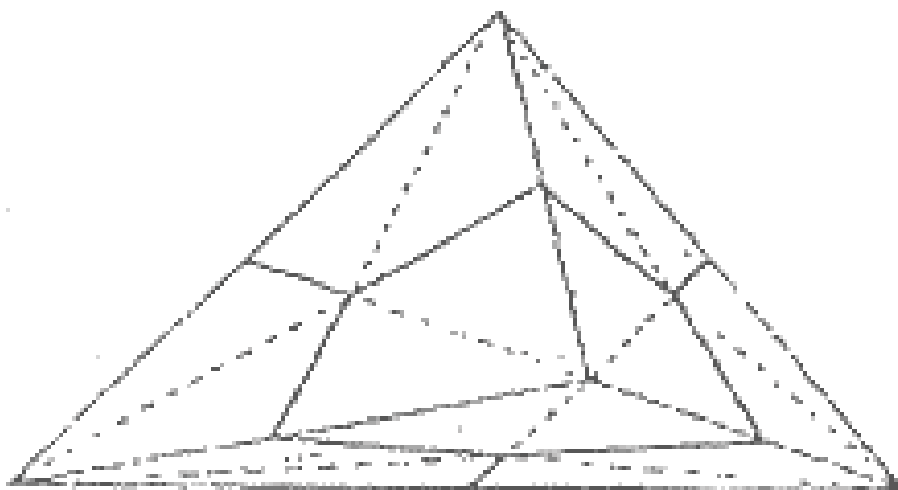
Les dispositifs de sollicitation et de recueil des réponses ne tiennent pas uniquement dans les questions, mais dans un ensemble de caractéristiques. Ainsi, selon que l'apprenant pourra répondre "à livres ouverts" ou non, une même réponse révélera des mécanismes mentaux (des compétences) tout(es) différent(es).

Nous comparerions volontiers le questionnement (ou le dispositif de sollicitation et de recueil des réponses) à un prisme qui est chargé de révéler des composantes particulières de compétences. L'observateur (à droite) reçoit une manifestation de la compétence (de la lumière) produite (à gauche).

Les schémas montrent que la même source lumineuse (le même phénomène cognitif) donne des observations différentes non seulement en fonction des grandes faces du prisme, mais de sa position ou de particularités de l'une ou de l'autre facette.

L'observation dépend donc non seulement de la forme du prisme, mais aussi de sa position (plus ou moins inclinée), l'ajout de miroirs, de lentilles filtrantes ou déformantes, de caches opaques, etc. Le tableau qui suit rend compte de ce prisme à plusieurs facettes (les colonnes) elles-mêmes ayant plusieurs côtés (les cases dans chaque colonne).

Le prisme est décomposé ci-après en plusieurs facettes sous forme de plusieurs colonnes juxtaposées. Chacune des facettes (PQRSTU) est séparée (indépendante) des autres, ce que veut illustrer le vide entre les colonnes. Les correspondances entre les lignes de colonnes différentes sont purement fortuites.



O. Objectifs <i>Compétences à évaluer</i>	P. Procédures <i>Conditions d'évaluation</i>	Q. Questions <i>Technique de sollicitation</i>	R. Réponses <i>Mode de recueil des observations</i>	S. Séquence <i>Liens des quest. entre elles et avec le test</i>	T. Training <i>et feedbacks formatifs</i>	U. Unités de tarif <i>et poids des questions</i>
6. Evaluation Jugement Prise de risque sous incertitude	5. On est réorienté si on s'enfonce dans une mauvaise voie	8. QROL Q à Réponses Ouvertes Longues	7. Répondre par une courbe (gaussienne) de probabilités déplaçable et rétrécissable ex : ELI (Van Lenthe)	6. Les questions sont conditionnelles au type d'erreurs (PACELBRO)	7. Diagnostic avec remédiation	7. Dépendant de la gravité de la réponse (ou de l'erreur)
5.Synthèse Expression Formulation	4. On peut poser des questions (demandes de précisions) sur la question	7. QROC courtes QROM multiples QROMO ordonnées QROMOP probabilisées	6. Distribution des probabilités sur toutes les solutions (avec Somme des p = 1)	5. Les questions sont liées entre elles (par exemple Double Check)	6.Commentaires sur sol. incorr. disponibles juste après la R	6. Dépendant du poids des questions ou de leur difficulté (théorique /statistiq)
4. Analyse Détection	3. On peut justifier son choix	6. QINT à intervalles QRIC = à choisir QROI ouvertes par intervalle	5. Intervalles avec taux prédéfinis de probabilité de "recouvrement" ex : Tertiles, fractiles et surprises (Pitz)	4. Les questions sont conditionnelles à la probabilité de réussite (Rasch)	5. Les Solutions Correctes sont disponibles juste après la réponse.	5. Dépendant de la rapidité de réponse
3. Application Transposition	2.Les questions sont à livres ouverts	5. QCM SGI (Solutions Générales Implicites)	4. probabilités continues a) numériques (ex : 34,21%) b) graphiques (ex : Confidence contour (Bruno)	3. Les questions sont conditionnelles à la probabilité de maîtrise/ non maîtrise (Wald)	4. Les scores totaux sont disponibles après chaque question	4. Dépendant des certitudes, et conformes à la théorie des décisions
2. Compréhension Interprétation	1. La modification des Réponses est permise (Double passage)	4. QCRM à réponses multiples (Vrai Faux généralisé ou VFG)	3. degrés-zones de certitude 0 (0% à 25%), 1(25-50%) 2(50 - 70%); 3(70 à 85%) 4(85 à 95%); 5(95 à 100%)	2. Les questions sont les mêmes pour tous ,mais l' ordre est au choix de l'étudiant	3. interactif (GUESS)	3. Correction for guessing classique
1. Rappel de mémoire évocation restitution		3. QCM simples : une solution correcte, parmi les proposées, une réponse à fournir	2. certitudes ordinales ex : peu sûr moyennement sûr très sûr	1. Les questions sont les mêmes pour tous et l'ordre est imposé	2. individuel (répertoires de questions publiées)	2. Ternaire (ex : -1, 0, +1)
0,5 Reconnaissance recognition		2. Vrai Faux omission permise	1. certitude binaire (Z=Zeker)		1. collectif (quizzes de fin de cours)	1. Binaire (ex : 0, 1)
		1. Détection (ex : repérer les fautes d'orthographe : omission = Faux	0. Sans commentaire (no comment)		0. aucun	

1. La facette P (les procédures)

(1) La modification des réponses

Dans un questionnaire classique (papier/crayon), l'étudiant dispose dès le départ de l'ensemble de questions, peut décider de répondre à certaines, répondre "au brouillon", etc. Cette possibilité est appréciée lorsqu'elle est aussi offerte sur ordinateur (pouvoir modifier sa réponse). Le logiciel CHECK permet, par exemple, que chaque question puisse être présentée deux fois; seule la réponse à la seconde fois est prise en compte.

(2) Les questions à livres ouverts

On vise à mesurer plus la compréhension et l'analyse que la connaissance (de mémoire). Les étudiants ne gagnent pas forcément du temps. Les livres avec index (des auteurs, des concepts) sont beaucoup plus faciles à consulter.

(3) La justification du choix

En même temps que la procédure des Solutions Générales Implicites, D. LECLERCQ (1993) a instauré la possibilité de justifier chaque réponse dans une ligne de "commentaires". Il est convenu que l'enseignant ne lira QUE les justifications qui se rapportent à des réponses incorrectes et, s'il est convaincu que le raisonnement est correct, mais que la réponse est mal "encodée", le professeur change la réponse en réponse correcte avant le traitement par ordinateur.

Le professeur en tire cependant la leçon pour sa propre question. Celle-ci est probablement passible d'amélioration puisque le professeur a été amené à valider une réponse incorrecte. On ne s'étonnera donc pas de ce que le professeur corrige en même temps les justifications ET ses propres questions.

Puisque le professeur a promis de ne pas lire les justifications des réponses correctes (manque de temps), l'étudiant n'a pas à craindre de perdre des points : les justifications ne peuvent lui être QUE bénéfiques. Il est vrai que lorsque le professeur lit (par inadvertance) des justifications de réponses correctes, il lui arrive de trouver de quoi considérer cette réponse comme suspecte ... mais il a promis de ne pas lire !

Les questions liées entre elles telles que Double Check (facette S = séquence) relèvent un peu de la même problématique : une réponse correcte est-elle preuve de compétence et une incorrecte preuve d'incompétence ?

(4) Poser des questions (demandes de précisions) sur la question

C'est ce qui est permis, en principe, par les examens oraux. On pourrait l'imaginer aussi pour les examens sur ordinateur (cliquer sur un mot pour obtenir des précisions sur ce mot SI LE PROFESSEUR LES Y A MISES).

(5) Réorientation si on s'enfonce dans une mauvaise voie

C'est aussi une caractéristique de l'examen oral. Le principe Double Check (facette S = Séquence) répond au même souci.

2. La facette Questions (techniques de sollicitation)

(1) Détection

La non-détection (ou omission) est assimilée à une erreur, de la même façon que la fausse alerte (détection abusive). Cette procédure s'applique à la relecture de textes en vue d'y détecter d'éventuelles erreurs d'orthographe.

(2) Vrai-Faux et omission

A côté de la confirmation (VRAI), de la dénonciation (FAUX), l'absence de réponse (OMISSION) est prévue.

Beaucoup y voient une façon d'avouer son doute et le préfèrent au choix forcé (où omettre équivaut à accepter par exemple).

De nombreux auteurs (1) considèrent que c'est une façon beaucoup trop peu subtile d'exprimer son doute et recommandent des systèmes de réponses (3) à (7) dans la facette "Réponses", systèmes où l'omission est même peu recommandée (et remplacée par "je réponds (VRAI ou FAUX) mais vous indique si c'est au hasard ou non).

(3) QCM simples : une solution correcte, une réponse

D. LECLERCQ (1986) a proposé une taxonomie des QCM présentant les dimensions Q, R, S, T, etc. où Q1R1 signifie "L'étudiant est averti que la question ne comporte qu'une seule solution correcte, qu'aucune solution "générale implicite" n'est d'application, et qu'il doit fournir une seule réponse.

(4) QCRM : QCM à Réponses Multiples

On est dans la situation Q2 R2 : "l'étudiant est averti que plusieurs solutions pourraient être correctes et qu'il peut donner plusieurs réponses, c'est-à-dire choisir plusieurs solutions comme étant correctes".

D. LECLERCQ (1986, p. 77) préfère parler de "Vrai-Faux Généralisé" (VFG) car il s'agit de répondre V ou F à chacune des solutions proposées, que la notation se fasse de façon additive (les points à chaque solution sont ajoutés; Ainsi 0, 1, 1, 0 font 2/4) ou multiplicative (les points à chaque solution sont multipliés; ainsi, 0, 1, 1, 0 font 0/1).

(5) QCM SGI : Solutions Générales Implicites

L'étudiant est averti qu'en plus des solutions (5 au maximum) dactylographiées, il doit prendre en considération 4 autres solutions, non rappelée :

6 = Rejet (aucune des k solutions présentées n'est correcte).

7 = Toutes.

8 = Manque de données pour que l'on puisse répondre (branches). La réponse correcte dépend d'une précision que l'énoncé ne comporte pas.

9 = Une absurdité s'est glissée dans l'énoncé (une contre vérité). Cette réponse est prioritaire sur toutes les autres.

Ces principes ont été décrits dans LECLERCQ (1986, pp. 127-144; 1992). BOXUS et al (1990).

(6) QINT : Questions à Réponses par Intervalles

Ce système n'est valable que pour les continuums (longueur, poids, nombre, durée, etc.). Il consiste à inviter l'étudiant à fournir deux valeurs, une limite inférieure (LI) et une limite supérieure (LS) entre lesquelles la valeur correcte doit se trouver. La notation (le scoring) peut se faire en tenant compte

- (a) de l'ampleur de l'intervalle : les points positifs (s'il inclut la RC) comme négatifs (s'il n'inclut pas la RC);
- (b) de la centration : les points positifs sont d'autant plus élevés que la RC est au centre de l'intervalle;
- (c) des deux : voir le système de LECLERCQ & TOBY (1995).

Plusieurs modalités de questionnement sont possibles :

- les QROI = par réponses ouvertes sous forme d'intervalle.
- les QRIC = par réponses sous formes d'intervalles à choisir.

(7) QROC = Questions à Réponses Ouvertes Courtes

C'est le cas lorsqu'on répond par un nombre, un mot, une expression (une définition) par exemple.

Les QROM consistent à accepter plusieurs réponses, et les points (totaux ou partiels) sont obtenus si l'une d'entre elle est correcte.

Ex. : Quelle est la capitale du Pérou ? Réponse : J'hésite entre Lima, Santiago et Quito.

Les QROMO sont des QROM où les réponses sont ordonnées/

Ex. : En premier lieu, Lima, ensuite Santiago enfin Quito.

Les QROMOP sont des QROMO avec probabilités réparties sur les diverses réponses.

Ex. : Lima (40 %), Santiago (35 %), Quito (25 %)

(8) QROL : Questions à Réponses Ouvertes Longues

C'est le cas des dissertations, rédactions, rapports de recherche, etc. On consultera le célèbre ouvrage "Evaluation continue et examens, Précis de docimologie de G. DE LANDSHEERE (1992, 6e édition) pour ses descriptions de phénomènes tels que l'effet Posthumus (tendance à répartir les notes selon une distribution gaussienne) et les effets de correction (de stéréotypie, de halo, de séquence...) étudiés par là le mouvement critique de la docimologie lancé par PIERON (1963).

3. La facette Réponses

(0) Sans nuance

La plupart du temps, pour les questions de type (1) à (7), on n'invite pas l'étudiant à commenter sa réponse, et c'est regrettable car il en est très capable, et ce "supplément méta-cognitif" est précieux à des fins de diagnostic et de remédiation, comme on le verra plus loin.

(1) Certitude binaire

L'étudiant indique s'il est ou non sûr de sa réponse. Des chercheurs hollandais invitent à entourer le Z de "Zekerheid" pour indiquer s'ils sont sûrs, et à ne pas l'entourer s'ils ne sont pas sûrs.

(2) Certitudes ordinales

Dans de très nombreuses études américaines, les étudiants ont indiqué s'ils sont "*totally sure, fairly sure, sure, weakly sure, not sure at all*" en ce qui concerne leur réponse. Le problème de ces repères ordinaux (et c'est aussi vrai pour les "binaires") est leur manque de précision. Pour certains étudiants, *fairly sure* c'est 20 % de chance, et pour d'autres c'est 80 % ! Le même **mot** utilisé par le même étudiant dans la même épreuve mais pour des questions différentes a parfois des significations différentes.

(3) Degrés-zones de certitude

Des zones ont été convenues avec des codes. Ainsi, on a beaucoup utilisé pendant les années 70, quatre zones : code

0	= entre 0 % et 25 %
1	= entre 25 % et 50 %
2	= entre 50 % et 75 %
3	= entre 75 % et 100 %

D. LECLERCQ (1983) a démontré expérimentalement que

- les estimateurs humains, adolescents et adultes sont capables de faire plus de nuances (sept environ) sur l'échelle des certitudes;
- ces nuances sont d'autant plus subtiles qu'on s'éloigne du centre de l'échelle (50 %) et que l'on se rapproche des extrêmes (0 % et 100 %).

Or, en évaluation pédagogique courante, on a peu besoin de faire des nuances dans la zone d'ignorance ou de faible connaissance (moins de 50 %), mais beaucoup plus dans la zone de connaissance (plus de 50 %) car l'apprentissage a assez souvent pour effet d'augmenter la confiance dans les réponses que l'on donnait déjà (correctement) auparavant, et pas seulement pour effet de changer les réponses.

Pour toutes ces raisons, D. LECLERCQ utilise les zones suivantes :

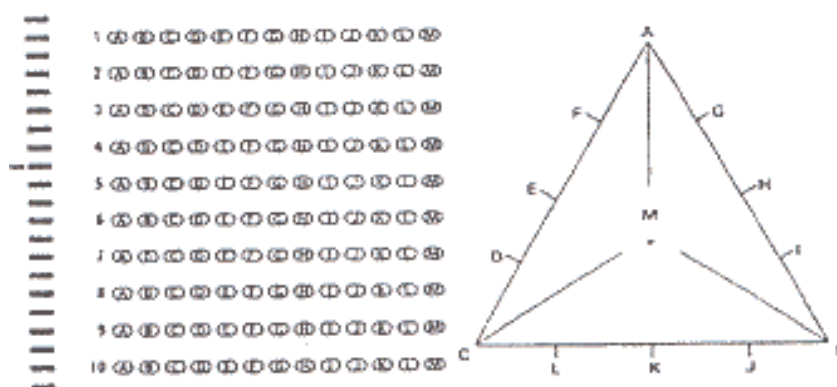
0	= 0 % à 25 %
1	= 25 % à 50 %
2	= 50 % à 70 %
3	= 70 % à 85 %
4	= 85 % à 95 %
5	= 95 % à 100 %

Chacune de ces zones comporte son "degré d'imprécision" (ou son "flou") puisque quelqu'un qui répond 2 peut tout aussi bien signifier "50 %" que "70 %" et, dans ces deux cas extrêmes, l'erreur (l'imprécision) de l'observateur est de 10 % puisqu'il considère que choisir le code de certitude 2 c'est être certain à 60 % (le centre de la zone).

(4) Probabilités continues

Fournir une probabilité (comprise entre 0 et 1) ou un pourcentage de chances (compris entre 0 et 100 %) sans aucune limite de précision (32,467 % de chances est autant accepté que la probabilité 0,189254). D. LECLERCQ (1983) a montré que même quand on leur en donne la possibilité, la très grande majorité des apprenants n'utilisent pas ces subtilités mais s'en tiennent aux grands points de repères tels que les dizaines (10 %, 20 %, 30 %, etc.) et des portions remarquables de l'unité (33 %, 66 %, 25 %, 75 %), soit une quinzaine de modalités de réponses.

DE FINETTI (1965) avait suggéré d'inviter l'apprenant à marquer d'une croix sur un triangle à quelle distance de trois solutions (situées aux trois sommets du triangle) est sa conviction quant à la réponse correcte. C'est ce principe qu'applique encore aujourd'hui BRUNO (1993) sous le nom américain de "confidence contour".



Cette modalité de recueil des certitudes est malheureusement limitée aux QCM à 3 solutions proposées, alors que les degrés de certitude sont valables pour toute forme de question (y compris les QROC à réponse ouverte courte).

(5) Intervalles avec taux prédéfinis de probabilité de recouvrement

Cette modalité est uniquement valable pour les réponses portant sur des valeurs numériques plaçables sur un continuum, par exemple, un nombre, ou une taille, un poids, une distance, etc.). L'étudiant est invité à estimer (approximer) la réponse correcte en avançant un intervalle (une limite inférieure et une limite supérieure) de telle sorte que cet intervalle recouvre la valeur réelle (exacte) dans x % des cas. (Exemple : Quelle est la longueur de cette table ? Réponse : entre 1,5 et 1,7 m).

PITZ (1974) appelle "surprises" les cas où la valeur réelle n'est pas comprise dans l'intervalle. Au début de l'utilisation de cette procédure, PITZ se faisait des illusions sur la précision des répondants et fixait à 90 % les chances de recouvrement (donc à 10 % la probabilité de surprise). En fait, les répondants obtenaient plus de 50 % de surprises. De guerre lasse, PITZ a fini par exiger 33 % de recouvrement seulement, avec 33 % de surprises vers le bas et 33 % de surprises vers le haut.

Couper ainsi les possibilités en 3 zones (grâce à 2 valeurs) a été appelé par PITZ la méthode des **tertiles**. Rien n'empêche cependant de fournir 4 valeurs (donc 5 zones) avec 33 % pour la zone centrale et 5 % pour les deux zones extrêmes (et donc 90 % pour les 3 centrales). Toutes ces consignes raffinées ont été appelées "**fractiles**" et préfigurent le système ELI de VAN LENTHE (voir modalité 7 ci-après).

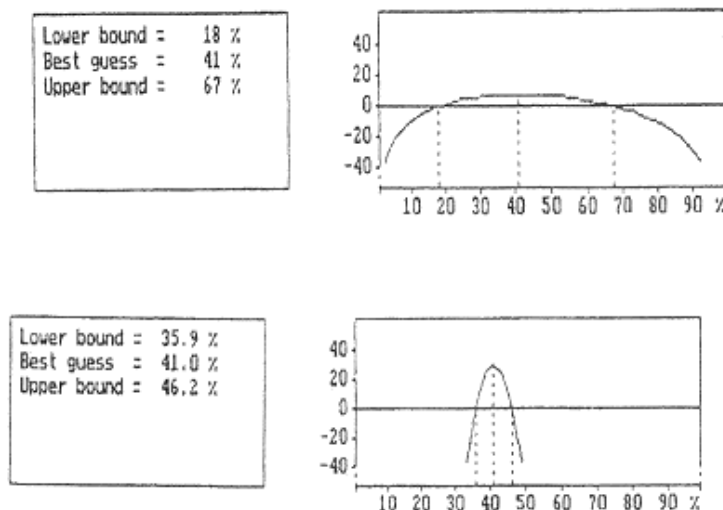
(6) Distribution des probabilités sur toutes les solutions

Dans cette modalité, d'habitude on garantit qu'il y a une et une seule réponse et solution correcte. L'étudiant doit répartir les 100 % de chances sur les diverses solutions proposées, la somme devant faire exactement 100 %. On voit en quoi cette technique permet d'observer les subtiles modifications intervenues après un apprentissage : même en cas d'octroi de la majorité des pourcentages (plus de 50 %) à la solution correcte, quelles sont les solutions incorrectes qui gardent néanmoins une certaine crédibilité, et de quelle intensité ?

D'habitude, on octroie comme points le pourcentage que l'étudiant a attribué à la réponse correcte, mais des scoring plus complexes existent (SHUFORD, 1966 et 1993: DIRKZWAGER, 1993).

(7) Répondre par une courbe de probabilités

VAN LENTHE, par son système ELI (comme "*ELICITATION of probabilities*") pose des questions dont les réponses peuvent être placées sur un continuum. Par exemple "Quel pourcentage des familles hollandaises possède deux téléviseurs ?" permet à l'étudiant de répondre en déplaçant horizontalement (sur l'axe des réponses possibles) une courbe (de gauche à droite sur l'axe des probabilités) et en la rétrécissant (diminution de l'écart-type) ou l'applatissant (augmentation de l'écart-type), et ce avec les 4 flèches du curseur.



La courbe représente d'ailleurs les scores (en ordonnée) que le répondeur obtiendra selon la position de la réponse correcte par rapport à cette courbe. On constate que le score est négatif quand la réponse correcte est en-deçà de l'endroit où la partie gauche de la courbe croise l'axe des abscisses. Il en va de même si la réponse correcte est au-delà de l'endroit où la partie droite de la courbe croise cet axe (VAN LENTHE, 1993).

4. La facette S : séquence

(1) L'ordre des questions est imposé

Une telle procédure peut avoir tout son sens si des questions ultérieures révèlent les réponses correctes aux premières (et qu'on empêche de modifier les réponses, par exemple, en retirant les feuilles des réponses à ces premières questions).

(2) L'ordre des questions est au choix de l'étudiant

C'est ce qui se passe le plus souvent dans des interrogations ou examens. L'étudiant a accès à toutes les questions et s'organise pour y répondre. La sagesse (ou l'habileté) dans la réponse aux tests (ce que les Anglo-saxons appellent "*test wiseness*") recommande de répondre d'abord aux questions pour lesquelles on voit clairement et sûrement la réponse correcte, puis de revenir aux autres questions, ce qui constitue une saine gestion du temps.

Dans une première version du logiciel CHECK (interrogation interactive par QCM sur ordinateur), l'ordre des questions était imposé, ce que les étudiants ont trouvé regrettable. Le logiciel a alors été modifié pour permettre un double passage; lors du premier, les réponses et certitudes étaient "provisoires".

(3) Les questions sont conditionnelles à la probabilité de maîtrise de la matière

En 1943, WALD mit au point une méthode statistique permettant, à partir de tirages successifs de pièces dans un lot (de 1000 pièces par exemple), et en comptant les pièces parfaites (1) et les défectueuses (0), de calculer la probabilité que le lot atteigne un taux donné (1 % par exemple) de pièces défectueuses. On considère les questions posées à un examen (oral par exemple) comme des tirages dans le lot des questions possibles. Au début des tirages, on dispose de trop peu de données pour décider dans un sens ou dans un autre. A partir d'un certain nombre de tirages et donc d'observations 1 ou 0, les formules mathématiques de WALD permettent de garantir que le risque de maîtrise (moins de 1 % de défauts) peut être accepté avec une probabilité suffisante (par exemple 95 % de chances de ne pas se tromper) ou qu'au contraire, étant donné la série d'erreurs constatées dans l'échantillon prélevé, on peut rejeter l'hypothèse de maîtrise. Tant qu'on n'est pas arrivé à l'une ou l'autre conclusion, le testeur continue à "prélever" (à poser des questions) et donc à mettre une note insuffisante à l'étudiant (MORONEY, 1978).

(4) Les questions sont conditionnelles à la probabilité de réussite de la question

En exprimant sur une même échelle (d'habitude l'échelle des WITS) la Compétence (C) d'une personne et la Difficulté (D) d'une question, il est possible, à partir de la formule mathématique de RASCH (1960) de calculer la probabilité qu'a cette personne de réussir cette question.

Au début du testing, on dispose d'une estimation de C (grâce aux réponses de l'étudiant à 10 questions calibrées, C étant la moyenne des D des questions auxquelles il a pu répondre correctement). Comme on sait par ailleurs que ce sont les questions qui ont une probabilité de succès de 0,5 qui ont le plus grand pouvoir informatif sur la compétence de la personne, les questions successives sont choisies dans la banque de manière à ce que $D = C$ car, dans le modèle de RASCH, quand $D = C$, alors p de succès = 0,5 (LECLERCQ, 1987, 29-41).

La compétence de l'étudiant est recalculée après chaque réponse, donc après chaque succès ou échec à une question de difficulté D connue.

(5) Les questions sont liées entre elles

les examens oraux fournissent de nombreux exemples de cette situation. Le professeur demande de décrire un phénomène, puis de l'illustrer, puis de le justifier, puis de l'appliquer, puis de le critiquer, etc. Le principe de la banque de questions DOUBLE CHECK consiste à présenter les QCM (SGI) en deux sous-questions : une PRIM et une BIS. Ainsi, le feed-back à la question PRIM (ex. : "L'énoncé de la question contenait une absurdité") se poursuit par une question appelant à préciser (ex. : En quoi consiste cette absurdité ?"; suivent alors plusieurs solutions proposées). On peut ainsi distinguer 4 situations selon que les réponses sont correctes ou non : situation de compétence et de vigilance (1,1), de compétence partielle (1,0), d'ignorance ou de non compréhension (0,0), de compréhension mais de non vigilance ou d'inattention (0,1).

Lors d'une récente expérimentation auprès des étudiants de l'agrégation, MAGAIN (1995) a montré que, lors des questions PRIM, les taux d'exactitude et d'erreur étaient respectivement de 65 % et 35 %.

Cependant, lors des questions BIS, ces taux se "subdivisent " comme suit :

Réussite lors de la PRIM 65 %	Réussite lors de la BIS 50 %	Compétence parfaite
	Echec BIS 15 %	Compétence partielle
Echec lors dela PRIM 35 %	Réussite BIS 20 %	Non vigilance
	Echec BIS 15 %	Incompétence

(6) Les questions sont conditionnelles au type d'erreur

Lorsque l'on dispose d'une typologie d'erreurs et de diagnostics précis quant à leur cause (en 1960, BONBOIR avait établi un tel dossier pour l'arithmétique à l'école primaire), on peut concevoir un système de remédiation conditionnelle. C'est le cas du PACELBRO (Programme Auto Correctif à Embranchements sous forme de livre brouillé à Réponses Ouvertes).

5. La facette T - training

(1) Entraînement collectif par quizzes de fin de cours

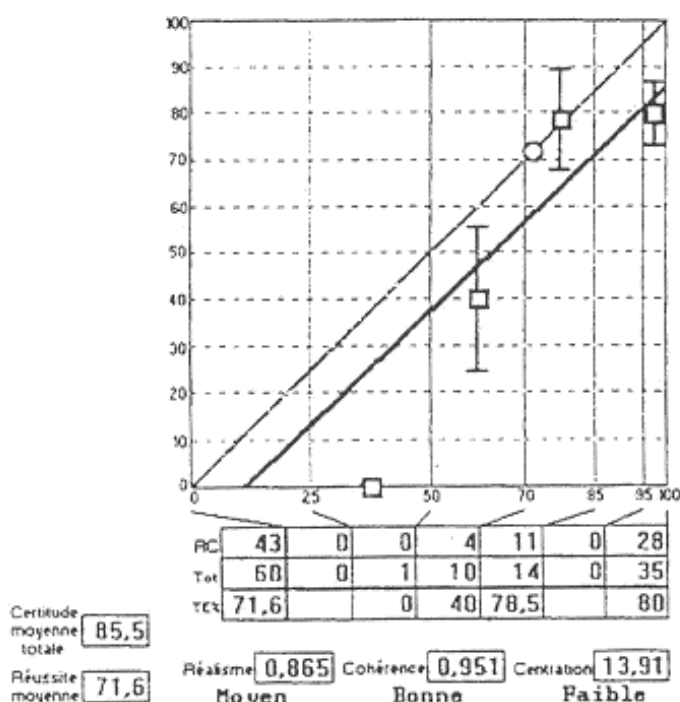
Nous appelons quizzes des séries de quelques QCM (5 ou 6) posées en fin de cours à de pures fins d'exercisation et collectivement. Les étudiants peuvent communiquer entre eux, les scores n'entraînent aucune sanction et les réponses correctes sont communiquées séance tenante. Un dossier individuel est remis à chaque étudiant avec ses résultats le lendemain ou la semaine suivante.

(2) Entraînement individuel sur répertoires de questions publiées

Il existe de nombreux ouvrages présentant les questions d'examens passés. Leurs réponses correctes sont disponibles pour permettre l'auto-administration et l'auto-correction. C'est le cas aux Etats-Unis où des tests tels que le Scholastic Aptitude Test (SAT) ou le Test of English as a Foreign language (TOEFL) sont disponibles en librairie dès qu'ils ont été administrés. Il en existe une nouvelle version chaque année. A l'Université de Liège, les professeurs VIGNERON, HOUSIER, LECLERCQ mettent de tels recueils en circulation.

(3) Entraînement interactif à l'autoestimation

Répondre à des questions sur ordinateur, avec feed-back immédiat constitue une forme d'entraînement particulièrement efficace. Certains types de performances peuvent être entraînées spécifiquement. Ainsi, le logiciel GUESS (LECLERCQ & GILLES, 1994; GILLES, 1995) permet de s'entraîner à l'utilisation des degrés de certitude, et ce sur des centaines voire des milliers de questions, qui consistent à deviner les lettres successives d'un texte. Le graphique de réalisme est dressé en temps réel.



(4) Entraînement interactif aux évaluations à livres ouverts.

Les étudiants de la FAPSE-ULG ont également la possibilité de se familiariser avec les modalités d'un examen à livres ouverts où les questions sont posées via l'écran d'un ordinateur. Ils emploient alors, dans une situation d'entraînement sans enjeu en termes de points, le logiciel WINCHECK utilisé lors d'examens interactifs sur ordinateur.

The screenshot shows the WINCHECK software interface. At the top, a title bar reads "Test". Below it, a status bar displays the file path "...afeim95\cafeim95.003", the student code "Code étudiant: 191159", and the current score and time: "0/0 [0.0/20]" and "29:42 (10)". Below the status bar, there are five buttons labeled ".1", ".2", ".3", ".4", and ".5". The main area of the window contains a text box with the following text: "Célestin FREINET, dans 'Les plans de travail' (1962) reconnaît lui-même que la FIXATION (le lundi) des plans de travail fait perdre un temps considérable, qu'on ne peut pas vraiment rattraper. Cet inconvénient est évité dans le système 1: Winnetka de WASHBURNE 2: Dalton de PARKHURST 3: Personalized System of Instruction de F. KELLER". Below the text box, there is a row of buttons for the answer: "Réponse: ?" followed by buttons "1", "2", "3", "6: Aucune", "7: Toutes", "8: Manque", and "9: Absurde". At the bottom, there is a "Fin du test" button, a "Certitude: ?" button, and a progress bar showing percentages from 0% to 100% with corresponding numbers 0 through 5.

WINCHECK permet de recourir

- aux QCM SGI (facette Q, option 5) ainsi qu'aux QCM simples et VRAI FAUX (qui peut le plus peut le moins).
- aux degrés de certitude (facette R, option 3).
- aux questions doubles : PRIM et BIS (facette S, option 5). L'exemple ci-dessus ne présente que des questions simples, ce que l'on détecte au fait que (en haut à gauche) les boutons des (5) questions sont précédés d'un point (PRIM) mais non suivis du point indiquant qu'il existe une question BIS.
- au "double passage" : comme lors des testings "papier-crayon" en auditoire, on peut visionner toutes les questions du test (et y répondre provisoirement) avant de répondre définitivement.
- de voir le temps restant s'écouler si l'on veut (case à cocher à droite en haut).
- de voir son score à la question et son score à l'épreuve (sur 20) si l'on veut (case à cocher au centre en haut).

N.B. Dans l'exemple ci-dessus, la réponse correcte est 9 "ABSURDITE". (Pour FREINET, le temps nécessaire à la fixation des plans de travail n'est PAS perdu.)

6. La facette U: unités de tarification

Le regroupement de ces unités sera appelé barème.

(1) Barème binaire

On ne fait pas la distinction entre échec et omission (voir Q1 = détection) qui reçoivent le même tarif (0 point), la réponse correcte recevant 1 point.

(2) Barème ternaire

Par exemple, la réponse correcte reçoit un point, l'omission reçoit 0 points, et l'erreur un tarif négatif (par exemple -1).

(3) La correction for guessing classique

Ce barème comporte un tarif en cas de réponse incorrecte (TI) calculé de telle façon que, dans une épreuve constituée de QCMs où parmi les k solutions proposées une seule est correcte, un étudiant qui répondrait à toutes les questions au hasard (probabilité de succès = $p = 1/k$), obtiendrait un score nul (0) au total de l'épreuve. Si TC (le tarif en cas de réponse correcte) vaut 1, alors TI vaut $-1/(k-1)$, donc -1 en cas de Vrai Faux, -0,5 avec 3 solutions, -0,33 avec 4 solutions, -0,25 avec 5, -0,20 avec 6, etc.

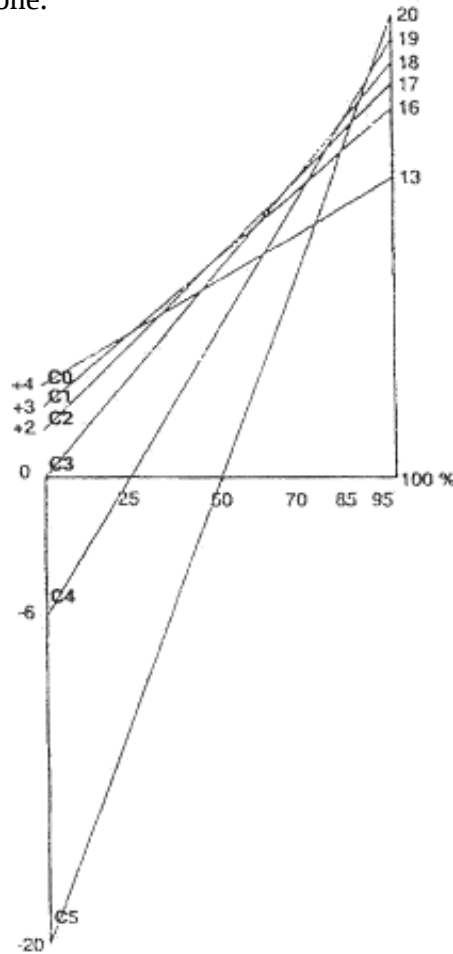
Cette procédure a été sévèrement critiquée (LECLERCQ, 1987, 1993; DELTOUR, 1995).

(4) Les barèmes dépendent des certitudes

Quand ils ont été calculés conformément à la théorie des décisions, ou, selon des "*proper scoring rules*", ces barèmes garantissent à l'étudiant qu'il a intérêt à exprimer son intime conviction sans la biaiser, bref à dire la vérité. Voici le barème des tarifs pour l'échelle présentée en R3 :

	Zones de %	TC	TI
Cert 0	0 - 25 %	13	4
Cert 1	25 - 50 %	16	3
Cert 2	50 - 70 %	17	2
Cert 3	70 - 85 %	18	0
Cert 4	85 - 95 %	19	- 6
Cert 5	95- 100 %	20	-20

Le graphique suivant garantit que chaque degré de certitude "optimal" pour la zone de certitudes annoncées : la droite des scores probabilisés surplombe toutes les autres (est maximale) dans cette zone.



(5) Barèmes dépendant de la rapidité de la réponse

Dans certains cas, il est justifié de tenir compte de la vitesse de réponse pour noter la performance. Par exemple, en interprétariat quand il s'agit de fournir un mot en traduisant un autre. Ou, en réanimation cardio-pulmonaire, quand il s'agit de vérifier (par les 3 critères) l'état de mort apparente. Ou en calcul mental, ou en chirurgie ou en sport, etc.

(6) Barèmes dépendant de la difficulté des questions

On peut imaginer donner plus de points à une personne qui répond correctement à une question difficile plutôt qu'à une question facile... et à l'inverse le pénaliser plus en cas d'erreur à une question facile. C'est ce qu'a fait CHERNOFF (1968) et ce que fait J. DELTOUR à Gembloux. Le problème est que, dans les deux cas, la difficulté est estimée uniquement sur base normative, c'est-à-dire sur les statistiques du groupe. Ainsi, il suffirait qu'une série d'étudiants comparses viennent rendre des copies nulles pour augmenter la difficulté statistique des questions¹. Il conviendrait d'estimer la difficulté sur base de raisonnements liés au contenu et aux processus mentaux.

* *
*

¹ C'est arrivé à J. DELTOUR qui, en conséquence, a décidé de ne pas tenir compte des copies nulles.

D. CONCLUSIONS

Trier, classer, relier les multiples variables impliquées dans les modalités de questionnements sont des démarches essentielles pour rompre avec une docimologie négative née de la contestation des examens traditionnels. Le kaléidoscope des techniques de questionnement constitue une tentative de clarification des facettes mises en oeuvre lors des situations d'examens.

Chaque facette révèle un large éventail de modalités. Ainsi, la facette « Q - Question » montre bien la place toute relative des QCM parmi les autres outils au service de l'évaluateur. Il convient d'être au clair en ce qui concerne les qualités et défaut de ceux-ci en vue d'opérer des choix judicieux. D. LECLERCQ (1986) a décrit en détail les avantages (brièveté de la réponse, simplicité de la correction, capacité à couvrir une matière, objectivité de la correction, ...) ainsi que les inconvénients (impuissance à mesurer la créativité, incapacité à entraîner l'étudiant à rédiger ses réponses, contraction du champs cognitif, ...) qu'offrent les QCM.

Les facettes peuvent, le cas échéant, être augmentées d'autres contenus que ceux présentés. Par exemple, il existe d'autres taxonomies d'objectifs, celle de GUILFORD (1967) ou de D'HAINAUT (1977) sont tout aussi valables quoique moins répandues.

D'un point de vue plus général, l'analyse de la combinaison des différentes facettes contribue à l'évaluation de la cohérence d'un dispositif d'examen. Ainsi, déclarer évaluer l'esprit critique des étudiants tout en employant que des QCM simples semble hasardeux si l'on se prive des ressources offertes par l'adjonction de solutions générales implicites qui elles permettent de tester la vigilance cognitive. De même, proposer les degrés de certitude aux étudiants sans que ceux-ci aient eu l'occasion de s'entraîner à leur utilisation nous paraît peu recommandable d'un point de vue déontologique.

A l'Université de Liège, dans le cadre des évaluations en candidature de la Faculté de Psychologie et des Sciences de l'Education, nous avons opté pour un dispositif ayant recours aux QCM-SGI avec degrés de certitudes pour évaluer des objectifs de type reconnaissance, compréhension, application et analyse-détection. Dans le cadre de certains cours les étudiants sont autorisés à consulter leur syllabus ainsi qu'à justifier leurs réponses. Un entraînement à ce type d'évaluation est proposé lors de quizzes en fin de cours ainsi que lors de trainings interactifs à l'aide de deux logiciels : GUESS pour l'entraînement à l'utilisation des degrés de certitude et WINCHECK pour la familiarisation aux modalités d'un examen informatisé à livre ouvert avec QCM-SGI et degrés de certitude.

Enfin, certaines facettes font apparaître des modalités qui nous paraissent peu recommandables : Q3 (QCM simples), R0 à R2 (absence de certitude ou certitudes ordinales), U1 à U3 (tarifs grossiers ou *correction for guessing* classique), T0 (absence d'entraînement). A la lecture des systèmes que nous avons nous mêmes adoptés (*cf.* WINCHECK), on comprendra les options et compromis (entre pertinence et faisabilité) que nous recommandons.

E. BIBLIOGRAPHIE

- BLOOM, B. *et al.* (1956). *Taxonomy of educational objectives. Handbook 1 : Cognitive domain*, New-York, McKay, traduit par M. Lavallée sous le titre « *Taxonomie des objectifs pédagogiques* », Montréal, Education nouvelle (1969).
- BOXUS, E. *et al.* (1991). *Principes communs pour évaluer les résultats cognitifs de la formation*, Bruxelles: Commission des Communautés européennes, EUROTECNET.
- BRUNO, J. (1993). Using testing to provide feedback to support instruction: a reexamination of the role of assessment in educational organizations. NATO ASI Series, *Item Banking: Interactive Testing and Self Assessment*, Berlin: Springer Verlag, 1993, Vol. 112, pp. 190-209.
- CHERNOFF, H. (1962). The scoring of multiple choice questionnaire, *Annals of Mathematical Statistics*, 33, 375-393.
- DE LANDSHEERE, G. (1992). *Evaluation continue et examens, Précis de docimologie*, Bruxelles: Labor.
- DE FINETTI, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item, *British Journal of Mathematical and Statistical Psychology*, 18, pp. 87-123.
- DELTOUR, J. (1995). Un système d'évaluation et de notation à base de QCM et de questions ouvertes, Journée FNRS, Systèmes de notations : choix et analyse statistique, 14 avril 1995, Gembloux.
- D'HAINAUT, L. (1977). *Des fins aux objectifs de l'éducation*, Ed. Labor, Bruxelles, 1977.
- DIRKZWAGER, A. (1993). A computer environment to develop valid and realistic predictions and self-assessment of knowledge with personal probabilities, in D. LECLERCQ & J. BRUNO, *Item Banking: Interactive Testing and Self Assessment*, Berlin: Springer Verlag, NATO ASI Series, 1993, Vol. 112, pp. 146-166.
- GILLES, J.-L. (1995). Entraînement à l'auto-évaluation : une comparaison filles/garçons à l'université, Actes du colloque « Enseignement supérieur : stratégies d'enseignement appropriées », Association Internationale de Pédagogie Universitaire, août 1995, Université du Québec à Hull, Canada.
- GUILFORD, J.P. (1967). *The nature of human intelligence*, New-York, Mc Graw-Hill.
- LECLERCQ, D. (1983). Confidence marking, its use in testing, 106 p. In POSTLETHWAITE & CHOPPIN, *Evaluation in Education*, vol. 6, 161-287, Oxford, Pergamon Press.
- LECLERCQ, D. (1986). *La conception des questions à choix multiple*, Ed. Labor, Bruxelles, 1987.
- LECLERCQ, D., BOXUS, E., DE BROGNIEZ, Ph., WUIDAR, H. & LAMBERT, F., (1993). The Taste approach: General implicit solutions in MCQ, open books exams and interactive testing and self-assessment, in D. LECLERCQ & J. BRUNO, *Item Banking: Interactive Testing and Self Assessment*, NATO ASI Series, 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. (1991). Peut-on utiliser l'auto-estimation à l'école ? *Enjeux*, 22, 57-63.
- LECLERCQ, D. (1991). Aider l'apprenant à détecter ses propres erreurs. *Enjeux*, 22, 64-73.
- MAGAIN, C. (1995). Banque de questions DOUBLE CHECK sur le cours de Psychologie Educationnelle à l'Agrégation, mémoire de licence, Université de Liège : Faculté de Psychologie et des Sciences de l'Education.
- MORONEY, M. (1978). *Facts from Figures*, New York : Penguin Book.

- PIERON, H. (1963). *Examens et docimologie*, Paris, P.U.F.
- PITZ, G. F. (1974). Subjective probability distributions for imperfectly known quantities. In Greg, L.W. (ed.) *Knowledge and Cognition*. New York: Adison Wiley, pp. 29-41.
- RASCH, G. (1960). Probabilistic models for some intelligence and attainment tests, Copenhagen : Danmark Paedagogik Institut.
- SHUFORD, E. *et al.* (1966). Admissible probability measurement procedures. *Psychometrika* 31, 125-145.
- SHUFORD, E. (1993). In Pursuit of the Fallacy: Ressurecting the Penalty, NATO ASI Series, *Item Banking: Interactive Testing and Self Assessment*, Berlin: Springer Verlag, 1993, Vol. 112.
- VAN LENTHE, J. (1993). The development and evaluation of ELI, an interactive elicitation technique for subjective probability distributions. NATO ASI Series, *Item Banking: Interactive Testing and Self Assessment*, Berlin: Springer Verlag, Vol. 112.
- WALD, A. (1943). *Sequential Analysis of Statistical Data*. New York : Columbia University Press.