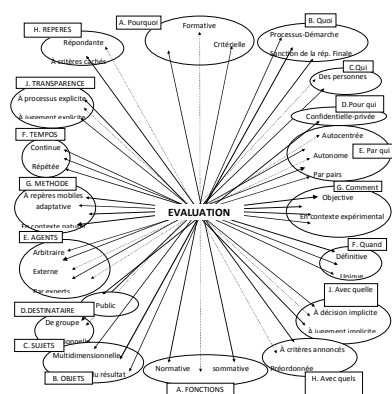
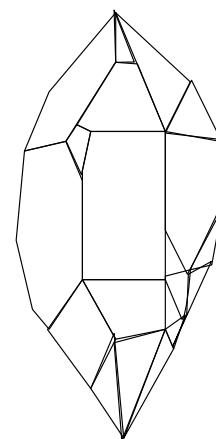


# Evaluation et Docimologie pour praticiens chercheurs

Editions de l'Université de Liège  
Chapitre 1



**Un PRISME ou  
une Rose des vents  
des  
Finalités et  
Caractéristiques  
des  
Dispositifs d'Evaluation  
des Acquis  
  
DEA**



Une rose des vents de l'évaluation... à en perdre la boussole ?

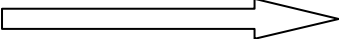
Introduction : à en perdre la boussole ? .....	2
A. Les <b>fonctions</b> de l'évaluation : POURQUOI ? .....	4
B. Les <b>objets</b> de l'évaluation : QUOI ? .....	8
B1. Produits vs Processus .....	8
B2. à <b>dimension</b> Unique vs Multiple.....	10
C. Les <b>sujets</b> de l'évaluation : QUI ?.....	18
D. Les <b>destinataires</b> de l'évaluation : POUR QUI ?.....	23
E. Les <b>agents</b> de l'évaluation : PAR QUI ?.....	24
F. Les <b>tempos</b> de l'évaluation : QUAND ? .....	28
G. Les méthodes d' <b>évaluation</b> : COMMENT ? .....	31
H. L'annonce des critères de l'évaluation : quelle PREVISIBILITE ?.....	36
I. La <b>transparence</b> de l'évaluation : avec quelle VISIBILITE ?.....	37

**Deux métaphores.** Selon le **PRISME** à travers lequel on évalue (mesure et juge) une performance, les images (chiffrées ou qualitatives) qui en résultent peuvent varier fortement l'une de l'autre. Or un prisme a plusieurs dimensions. On peut aussi parler d'une **ROSE DES VENTS** qui aurait autant de (pôles) Nord et de (pôles) Sud qu'il y a de dimensions (ou directions), dont le sens est indiqué par autant d'aiguilles qu'il y a de dimensions, chacune étant représentée par un axe à double flèche indiquant les deux **pôles opposés**. Ces dimensions ou facettes sont **INDEPENDANTES** l'une de l'autre. Pour décrire un DEA, il faut donc les décrire toutes.

## Une rose des vents des fonctions et caractéristiques de l'évaluation...à en perdre la boussole ?

### Un besoin de précisions conceptuelles

En 1972, Bloom, Krathwohl et Masia ont popularisé l'opposition entre « évaluation sommative » et « évaluation formative » les termes même du titre de leur « *Handbook of formative and summative evaluation* ». Or, au cours de ces trente dernières années, la recherche et les pratiques en évaluation se sont considérablement développées et le besoin de précisions conceptuelles et terminologiques) a lui aussi grandi.

Illustrons ce propos par un exemple (ci-contre) : l'évaluation d'un travail de fin d'études universitaires, portant sur deux composantes : un document écrit d'une centaine de pages et une défense orale devant un jury de trois personnes et un éventuel public. Nous invitons le lecteur à lire cet exemple maintenant. 

On le voit, sans parler du contenu (chapitre 2) ni des techniques (chapitres 3 et 4), une évaluation présente des **facettes** multiples (fonction, procédures, critères, etc.), susceptibles chacune de prendre des **dimensions** diverses (par exemple dans la facette « procédure », le caractère définitif ou révisable de la note est une telle dimension, le caractère unique ou répété de la performance en est une autre).

Ces dimensions sont relativement **indépendantes**. C'est pourquoi nous présentons ci-après une série de **dimensions bipolaires** regroupées (arbitrairement) dans des facettes dont le nom ne joue qu'un rôle de structuration du texte. Le nombre de dimensions (en général ou facette par facette du **prisme**) n'est pas déterminé et de nouvelles dimensions apparaissent, que l'on doit juger à chaque fois à leur pertinence dans la problématique évaluative considérée.

### Le mémoire et ses facettes d'évaluation

L'étudiant connaît sa note finale, les notes des trois membres du jury, leur note portant sur l'écrit et celle sur la défense orale, leurs commentaires oraux, voire écrits.

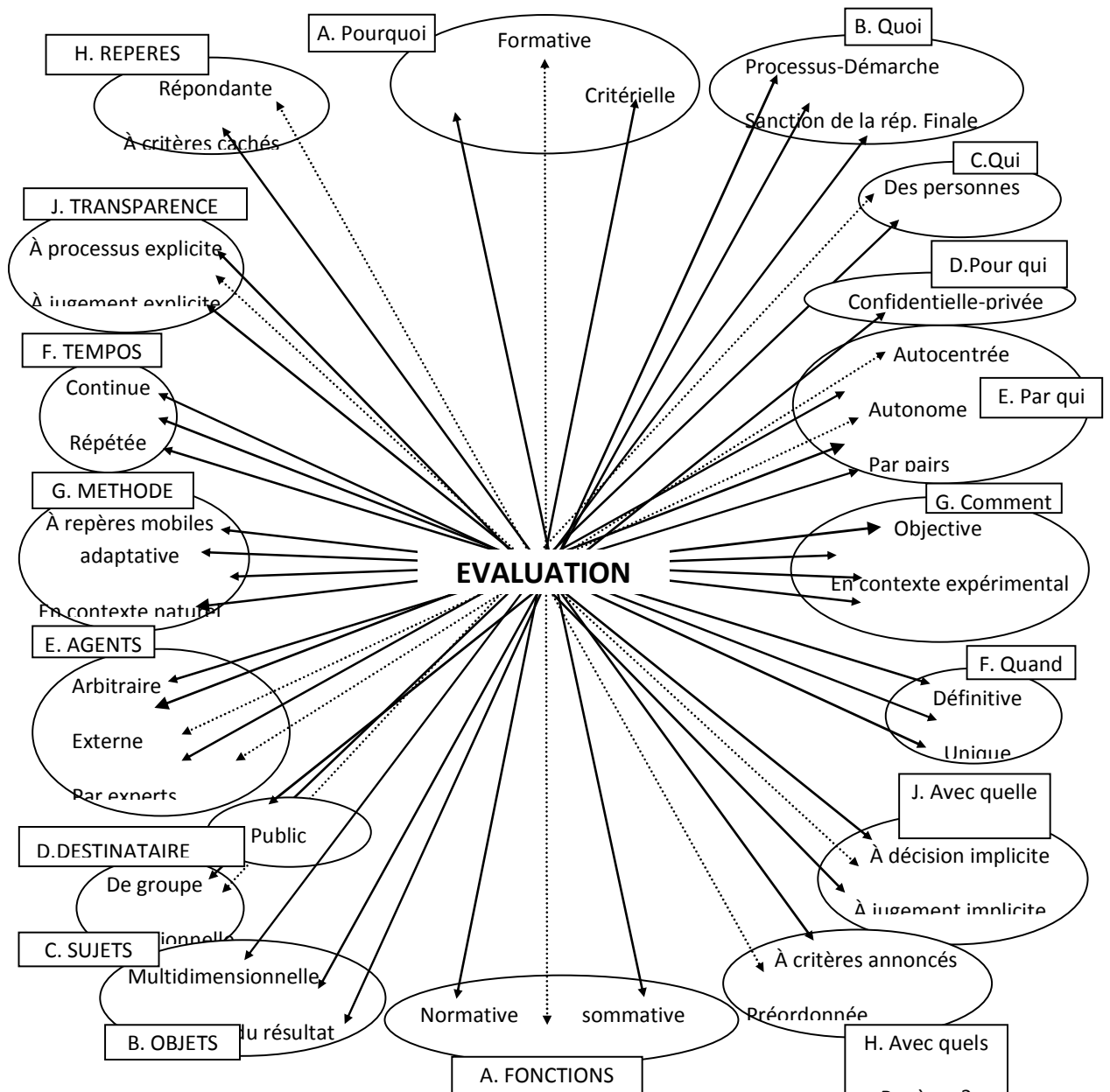
Cette évaluation a une **fonction** (ou un but, une visée) sommative, certificative et, à ce titre, est définitive (en tout cas pour la note), même si les commentaires ont un caractère formatif, mais pour la vie professionnelle future de l'étudiant, et non pour sa note, qu'il ne peut plus améliorer.

Il existe des **procédures** cependant où un tel travail peut être présenté en « dry run » (tournage à vide), avec des notes « provisoires » et la possibilité pour l'étudiant de soumettre à nouveau son travail remanié dans un certain délai. La nouvelle note peut soit « effacer » l'ancienne ou au contraire la compenser pour la moitié des points par exemple. On peut aussi imaginer qu'au cours des trois dernières années d'études chaque étudiant doit remettre un mémoire<sup>1</sup>, et que la note finale soit la moyenne des trois. Il y aurait ainsi répétition de cette performance qui, dans la plupart des cas est « unique » dans le curriculum de l'étudiant.

Sur les **critères**, les options de divers jurys peuvent être très différentes, certains ne considérant que le résultat final, d'autres prenant en compte le processus, notamment le degré d'aide fourni par les encadrants au mémorant.

<sup>1</sup> Cette procédure était d'application pour chacun des 9 cours de la Maîtrise en Pédagogie de la Santé, coordonnée par le Professeur J.F. d'IVERNOIS, à l'UFR de la Faculté de Médecine Leonard de Vinci (Bobigny, Hôpital Avicenne) de l'université de Paris Nord (Paris XIII)

La notion d'évaluation est « lisible » selon diverses **facettes**, décomposables en **dimensions** ou perspectives, directions, axes, souvent définis par deux pôles (ou deux sens) opposés.



On constatera qu'ici nous éviterons les « raccourcis de langage » fréquents, et utiles...mais à la condition impérative que l'on se comprenne !!!

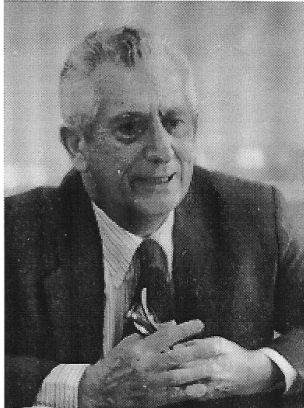
Ainsi, nous ne dirons pas « Evaluation formative » mais « Evaluation à visée formative », ni « Evaluation critérielle » mais « Evaluation à référence critérielle ».

On comprend dès lors mieux pourquoi, à notre sens, la question « cette évaluation est-elle formative ou critérielle ? » n'a pas de sens : ce sont deux dimensions différents qui ne s'excluent pas.

C'est un peu comme si l'on demandait « Cette dame est-elle grande ou suédoise ? » ?

## A. Les FONCTIONS de l'évaluation : POURQUOI ? ou POUR QUOI ?

### Un peu d'histoire



Prof. Benjamin Bloom (1913-1999) Université de Chicago. Photo à l'université de Liège en 1984

C'est avec leur célèbre ouvrage de 1971 « Handbook of formative and summative evaluation » que Bloom, Hastings et Madaus (Eds) ont lancé le concept d'évaluation formative (et, hélas) son opposition à « évaluation sommative ». Nous tenterons de montrer qu'il est plus pertinent d'opposer formative à sanctionnante (ou certificative) et sommative à diagnostique. Dans cette dernière opposition, il peut être utile d'associer (par jeu de mot) « sommative » à « sommaire », du moins dans son expression, parce qu'elle « résume » en un nombre (ex : 12/20), en un grade (ex : TB), quelque soit le nombre d'indices ainsi « sommés ».

Cet imposant ouvrage collectif fournit des exemples dans chacun des grands domaines scolaires non seulement d'objectifs inspirés des taxonomies d'objectifs cognitives (Bloom et al., 1956), affectifs (Krathwohl, Bloom et Masia, 1964), sensori-moteurs (Harrow, 1972), mais d'outils d'évaluation destinés à mesurer le degré d'atteinte ou de maîtrise de ces objectifs.

D'autres ouvrages s'évertuèrent à préciser les objectifs et les instruments d'évaluation. En français, retenons Vandeveldt et Vanderest (1975), V. et G. de Landsheere (1975) et D'Hainaut (1978).

L'évaluation formative est intimement liée à un autre concept bloomien : la pédagogie de la maîtrise (mastery learning).

### A1. Evaluation à but ou visée FORMATIVE vs SANCTIONNANTE

#### a) L'évaluation à visée formative

*« Evaluation intervenant, en principe, au terme de chaque tâche d'apprentissage et ayant pour objet d'informer élève et maître du degré de maîtrise atteint et éventuellement, de découvrir où et en quoi un élève éprouve des difficultés d'apprentissage, en vue de lui faire découvrir des stratégies qui lui permettent de progresser...L'expression « évaluation formative » due à Cronbach et à Scriven marque bien que l'évaluation fait ...partie intégrante du processus éducatif normal, les « erreurs » étant à considérer comme des moments dans ...l'apprentissage, et non comme des faiblesses répréhensibles ou des manifestations pathologiques. ...L'évaluation formative permet aussi de déterminer si un élève possède les prérequis.» (de Landsheere, 1979, 113).*

Cette expression servait au départ à désigner le processus d'amélioration de programmes (ou de curriculum). C'est Glaser (1963) qui en a appliqué le sens à l'apprentissage.

L'évaluation formative est au service de l'amélioration de l'apprentissage et ne fige pas la situation, dont elle facilite l'évolution par des remédiations. L'évaluation formative est, par définition, tournée vers le présent (et déborde sur le passé et le futur proches). Elle peut être auto-administrable et elle a le plus souvent un caractère privé (connu seulement de l'apprenant, de son professeur, et, éventuellement des parents), répétitif (autant de fois que nécessaire) et éphémère (le constat change d'une fois à l'autre). Elle est le plus souvent à référence critérielle (la performance et non « les autres étudiants »). Cependant, dans un but formatif, un entraîneur peut situer un sportif dans le classement mondial (ce qui est normatif).

**b) L'évaluation à visée sanctionnante** est au service d'une régulation extérieure à l'apprentissage (fonctionnement du système scolaire, recrutement dans une entreprise, etc.). L'évaluation peut être sanctionnante de deux grandes façons :

- la certification (tournée vers le passé);
- la sélection (tournée vers l'avenir).

## A2. Evaluation à référence NORMATIVE ou Relative vs CRITERIEE ou absolue

### a) L'évaluation à référence critérielle

Cette expression « Criterion Referenced Tests » est due à Robert Glaser, 1963.



Robert Glaser, Directeur du Learning Research & Development Centre (LRDC) à Pittsburgh en 1981.

Cette évaluation, propre aux EXAMENS, ne s'embarrasse pas des normes du groupe, mais considère des critères, des références absolues, fixes, en principe les objectifs à atteindre. Une question se justifie dans ces tests si l'objectif spécifique qu'elle représente le justifie, même si cette question n'est pas discriminante.

Le score minima à atteindre sera ici fixé quel que soit le taux de personnes qui l'atteindront. Ne pas connaître ce taux à l'avance expose l'examineur aussi bien à ce que TOUS réussissent ou à ce que TOUS échouent.

Pour éviter cette incertitude, nombre d'évaluateurs notent d'abord de façon critériée (et donnent à chacun les points qu'il mérite) puis transforment ces notes (ratings ou marks si elles sont chiffrées) de façon normative, et laissent passer les y % supérieurs ou les x meilleurs, ou ceux dont la note Z est supérieure à -1, etc.

Pratiquer de la sorte fait dépendre la réussite d'une personne de la performance des autres. Pour un candidat, il vaut mieux, dès lors, se présenter une année où les autres concurrents sont faibles.

### b) L'évaluation à référence normative

Cette évaluation exprime les résultats d'une personne X en termes de position dans les résultats d'un groupe, ces résultats servant de norme, de repère. Ainsi, on dira de X que son résultat est le 4e, ou supérieur à la moyenne, ou à 2 écarts-types au-dessus de la moyenne, ou de note  $Z = 2$ .

Cette évaluation est le propre des CONCOURS (sanctionnant par sélection) où les x « meilleurs » sont choisis ou réussissent, quelle que soit la valeur moyenne des candidats. Dans une telle approche, sont privilégiées les questions à haut pouvoir discriminant (donc ni trop faciles ni trop difficiles), l'idéal étant alors 50% de réussite à la question.

Le pouvoir discriminant d'une question ou des mesures résultant d'un test est expliqué dans Leclercq (1987,43-76), pages reproduites au chapitre 10 du présent ouvrage.

### c) Situations mixtes

#### C1. Evaluation mi –critérielle et mi-aléatoire

Par exemple tirer au sort parmi tous ceux qui ont obtenu des scores suffisants.

Ex : A l'entrée en faculté de médecine à Maastricht pour sélectionner les 200 entrants du numerus clausus.

#### C2. Evaluation mi –critérielle et mi-normative

Ex : Quand une classe du secondaire repère ses meilleurs candidats à l'émission de la RTBF « Génies en herbe », mais qu'elle ne peut en présenter que 4 sur le plateau et qu'elle a 9 candidats qui tous satisfont aux critères.

### A3. Evaluation à précision SOMMATIVE vs DIAGNOSTIQUE

#### a) L'évaluation sommative (ou globale)

Cette évaluation résume en un score (ex : 12/20) ou un niveau (ex : Distinction) un ensemble de sous-mesures.

C'est un bilan « sur un ensemble de tâches constituant un tout, correspondant, par exemple, à un chapitre de cours, à l'ensemble du cours d'un trimestre » (de Landsheere, 1979, p.115) et non au terme de chaque tâche d'apprentissage, ce qui l'oppose aussi à l'évaluation formative.

Ainsi, le grade de fin d'études universitaires (Satisfaction, Distinction, etc.) résume un ensemble de scores dans les différents cours et n'indique ni où le sujet excelle, ni ses lieux de faiblesse. A fortiori, il est muet (même à titre d'hypothèse) sur les causes éventuelles de ces faiblesses.

#### b) L'évaluation diagnostique (ou explicative)

Cette évaluation précise les lieux d'excellence et les faiblesses, détaille les processus erronés, éventuellement leur cause.

##### Exemple 1 :

Dans les **Questions « double face »**, la question BIS porte sur la justification de la réponse à la question PRIM, ce qui permet d'établir un diagnostic différentiel entre le manque de compréhension ou le manque d'analyse.

(voir Leclercq, Edumétrie et Docimologie, 2005, ch.4)

##### Exemple 2

Le logiciel (d'intelligence artificielle) **BUGGY** de J.S. Brown et R. Burton (1980) essaye de réaboutir à la solution erronée de l'évalué, par application de *bugs* successifs dans le raisonnement.

(voir Leclercq, 2005, Questions approfondies d'évaluation)



Anne Bonboir  
(UCL)

##### Exemple 3

Dans le système **PACELBRO** (Programme Auto-Correctif à Embranchement sous forme de Livre Brouillé à Réponses Ouvertes), après avoir fourni sa réponse (construite), l'évalué la compare à une série de réponses proposées (chacune représentative d'une erreur spécifique de raisonnement), ce qui lui permet de s'auto-diagnostiquer.

Cela a été permis grâce au travail d'enquête auprès de dizaines d'élèves par l'analyse de la réflexion parlée enregistrée pratiquée par A. Bonboir (1960) visant à identifier les démarches mentales erronées

Voici un extrait des (très nombreuses) observations de Bonboir :

Pour des exercices du type 6% de 200 FB, La		
réponse	Indique un(e)	
6 FB	Oubli de multiplier par le nombre de centaines	A
20 FB	Oubli de multiplier par le nombre de %	A
1200 FB	Oubli de diviser par 100	A
12	Oubli les unités	B
206 FB	Addition au lieu de prendre le pourcentage	C
14, 16 ou 18 FB	Erreur de table de multiplication	D

L'impact des remédiations a été décrit ailleurs (Leclercq, 2005, Conception d'Interventions et Construction de Produits éducatifs. Ed. Univ. de Liège.

NB : Nous ne considérons pas que fournir les solutions correctes aux étudiants après que ceux-ci aient répondu à une épreuve suffisée à en faire une évaluation diagnostique.

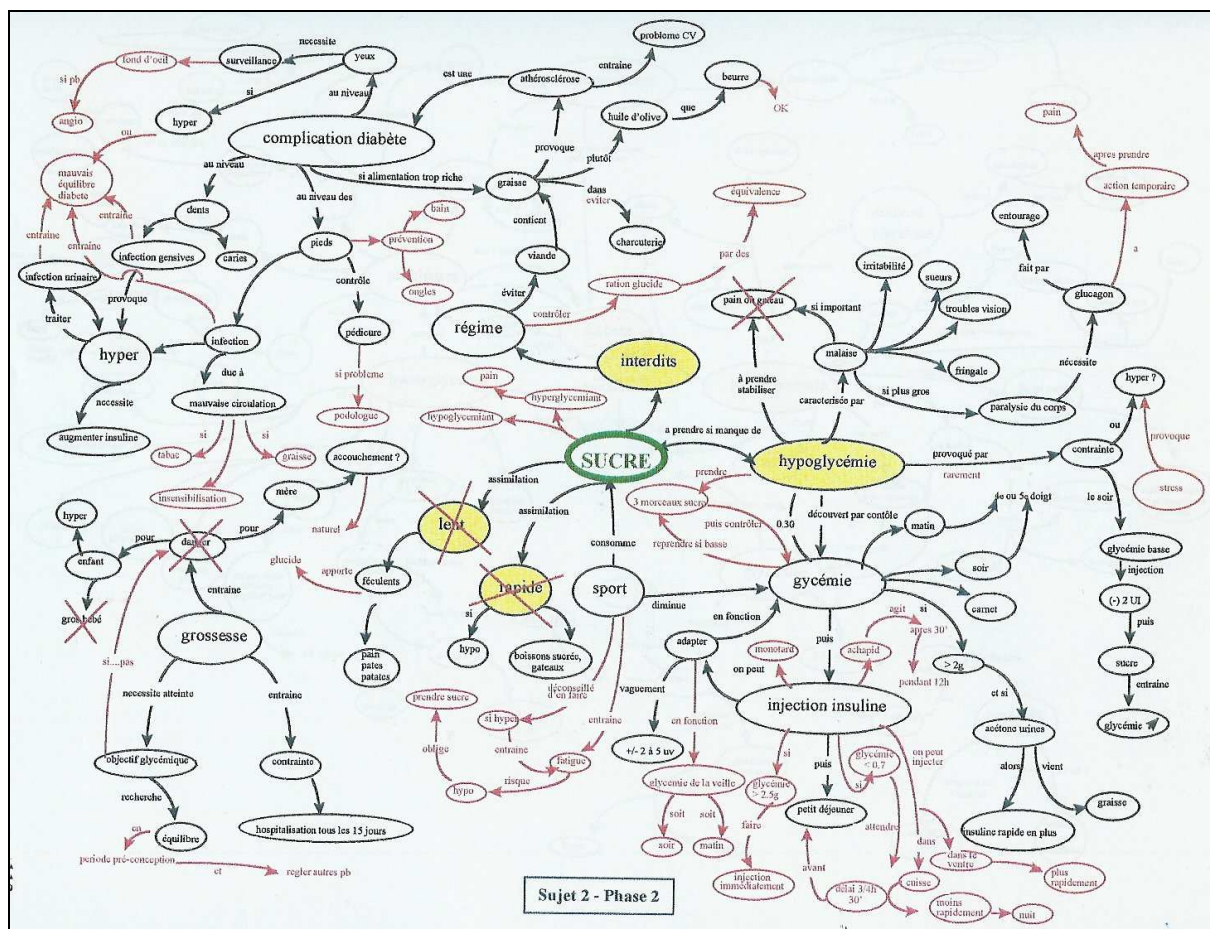


### Exemple 4 : La carte conceptuelle en éducation du patient diabétique par Claire Marchand (2001), chercheur à l'UER Pédagogie de la Santé. Léonard de Vinci. Université de Paris XIII.

Le principe classique de la carte conceptuelle consiste à demander à une personne d'écrire un mot-clé (par exemple « protéine ») au centre d'une feuille blanche, puis d'y « lier » tous les autres concepts qui, de son point de vue, s'y rattachent, en pré (parfois appelé « base ») et en post-test, c.-à-d. **avant et après une intervention**, parfois juste après (post-test **immédiat**) puis lors d'un deuxième post-test (**différé**).

Cette approche classique a plusieurs inconvénients. D'abord, le sujet n'écrit pas ses concepts au même endroit d'une fois à l'autre, ce qui rend la comparaison très difficile. Ensuite, on ne dispose pas de l'ordre dans lequel le sujet a écrit les concepts ; or l'ordre est un indice de leur disponibilité (saillance) mentale. Enfin, le sujet n'indique pas les liens entre les concepts : il les relie simplement par un trait, ce qui est difficilement interprétable car flou (comme on va le voir).

Dans la méthode de C. Marchand l'intervenant(e) écrit elle-même sur une (grande) feuille de papier (A3) ce que lui dit le patient. Comme dans l'entretien clinique Piagétien (1936), ou dans l'entretien d'explicitation de Vermeersch (1994), l'interrogatrice improvise les sous-questions de manière à rendre la carte complète. Par exemple, elle demande la signification de chaque trait (ou relation entre deux concepts) : « est un, a pour exemple, ne s'applique pas avec, ne se fait qu'en cas de ... ». On se doute que, dans une telle carte conceptuelle, les relations inter-concepts sont tout aussi importantes que les concepts eux-mêmes. En outre, lors du post-test, l'interrogatrice (ré)écrit (dans une autre couleur) sur la feuille du pré-test, rendant les modifications très visibles. Voici un exemple de carte conceptuelle (où le mot central imposé est « sucre »), en post-test immédiat, réalisée par Claire Marchand (2000,113), les nouveaux concepts en rouge et les « disparus » (par rapport au prétest) marqués d'une croix :



## B. Les OBJETS de l'évaluation : QUOI ?

### B1. Les PRODUITS ou Résultats vs les PROCESSUS ou Démarches

#### a) L'observation des résultats finaux

La phrase « Tout est bien qui finit bien » considère le résultat final, et non les cheminements nécessaires pour y arriver.

##### Exemple :

Pour entraîner à la compréhension de l'anglais parlé (Understanding Spoken English - USE), des logiciels comme EPEL et AUDIO-SCRIPT présentent à l'écran un texte où chaque lettre est remplacée par des pointillés.

.....  
.....  
.....

L'étudiant entend ensuite le texte :

The magical number seven.  
All my life I have been persecuted  
by an integer.

et est invité à introduire, via le clavier, les mots qu'il a compris à l'audition. Exemple :

The magical .....  
All my life. ....  
.....

Evaluation : On compte le nombre de mots remplacés après 1, 2 et 3 auditions.

#### b) L'observation des processus, des démarches

Les logiciels EPEL et AUDIO-SCRIPT permettent aussi à l'étudiant de demander de l'aide, par exemple, désigner un mot (ses pointillés) avec la souris

The magical .....  
.....

All my life . ....  
.....



puis obtenir (au choix) sa première lettre (P), ou sa dernière lettre (d) ou sa première partie (Per) ou sa fin (cuted) ou sa nature grammaticale (verbe au passé) ou, finalement, le mot entier (persecuted), puis sa traduction (persécuté).

Par l'enregistrement des traces, on peut observer les aides qui furent nécessaires pour aboutir à la réponse après 3 écoutes (la découverte de 6 mots sur les 14 à trouver).

##### Exemple 2 :

Pour savoir si l'élève est capable non seulement d'aboutir à la solution correcte d'un exercice mathématique, mais aussi d'y arriver par des démarches correctes, Halleux (1969) demande aux étudiants de **détecter des erreurs dans plusieurs démarches** (solutions d'une QCM) aboutissant pourtant toutes à la même réponse correcte, mais parfois par compensation de plusieurs erreurs (Leclercq, 1986, p. 77).

##### Exemple 3

Pour évaluer le niveau de conceptualisation algébrique atteint par les élèves, Réginald Burton (1999) observe les **modalités de résolution** (intuitives ou algorithmiques) pratiquées par des élèves résolvant des équations du premier degré à une inconnue.



### c). Exemple 4 : l'évaluation des enseignements universitaires

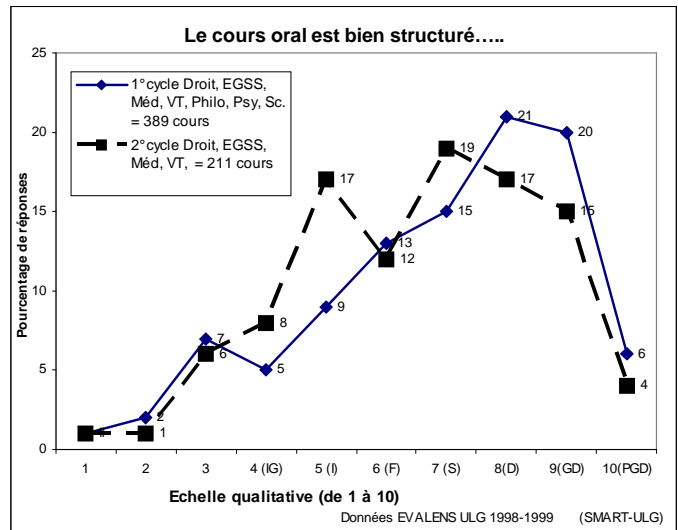
Un volet de l'évaluation des enseignements, l'avis des étudiants est réalisé via une enquête de satisfaction auprès des étudiants qui remplissent des questionnaires sur les cours qu'ils ont suivis. A l'université de Liège, le SMART aide les facultés à réaliser cette opération annuellement sur des centaines de cours.



#### 1°. Evaluation du processus

Très souvent, les questions posées portent sur « la façon dont cela s'est passé » :

- Le syllabus -s'il existe- était-il disponible à temps ? Son prix était-il raisonnable ?
- Le syllabus (ou le livre) -s'il existe- était-il clair ? utile ?
- Le professeur était-il ponctuel ? disponible en dehors des exposés ?
- Son débit oral était-il clair ? Son tempo était-il approprié ?
- Rendait-il l'auditoire actif ?
- Supportait-il ses exposés par des documents (audio visuels ou autres : témoignages)
- Les consignes de l'examen étaient-elles claires ? Et les critères de correction ?
- Le professeur a-t-il préparé à l'examen ?
- Les questions d'examen étaient-elles en concordance avec la matière ? sans ambiguïté ?
- Etc. On trouvera une synthèse de ce genre d'approche dans Huguette Bernard (1992).



#### 2°. Evaluation des produits ou impacts (déclarés et réalisés)

D'autres, plus rares (Bujold, 1997, 409-419 ; Rheau, 1995; Leclercq et Gilles, 1997) font porter leurs questions sur :	Quelle importance déclarée ont les objectifs suivants ?	Dans quelle mesure ont-ils été atteints selon vous ?
Donner le goût d'approfondir certains contenus	Nulla faible moy forte T forte	Nulla faible moy forte T forte
Tenir au courant des développements les plus récents de la matière	Nulla faible moy forte T forte	Nulla faible moy forte T forte
Faire comprendre les concepts fondamentaux en profondeur	Nulla faible moy forte T forte	Nulla faible moy forte T forte
Habituer à consulter la littérature scientifique et les bibliothèques	Nulla faible moy forte T forte	Nulla faible moy forte T forte

Que les professeurs et les étudiants ne donnent pas les mêmes réponses à ces mêmes questions ne surprendra pas. En outre, des processus très appréciés au service d'objectifs sans intérêts ou non atteints sont de peu de pertinence. Et l'inverse !

Beaucoup de curriculums ont dans leurs déclarations d'intention des objectifs du type de ceux qui sont énoncés ci-dessus...mais ne les atteignent pas, pour diverses raisons, entre autres parce que les processus (les méthodes) ne s'y prêtent pas. Il serait donc pertinent d'inverser l'ordre des questions : celles sur les objectifs et sur leur atteinte d'abord, puis celles sur les processus, par exemple en ajoutant une colonne à droite du tableau ci-dessus, avec l'intitulé « Les méthodes contribuaient-elles à l'atteinte de cet objectif ? ». On sait que de grandes réformes pédagogiques telles que l'APP (Apprentissage Par Problèmes) ou PBL (Problem Based Learning) visent à atteindre des objectifs ambitieux, via des méthodes qui leur sont cohérentes et appréciées par les étudiants.

Voir des Principes sur l'évaluation anonyme de l'enseignement par les étudiants à [www.caut.ca/français/ausujet/principes/questionnaires.asp](http://www.caut.ca/français/ausujet/principes/questionnaires.asp)

## B2. Evaluation à dimension UNIQUE ou à dimensions MULTIPLES

La plupart du temps, on se contente de l'exactitude de la réponse, indépendamment du temps mis pour la produire, des aides qui furent nécessaires, de la certitude qui l'accompagne, etc.

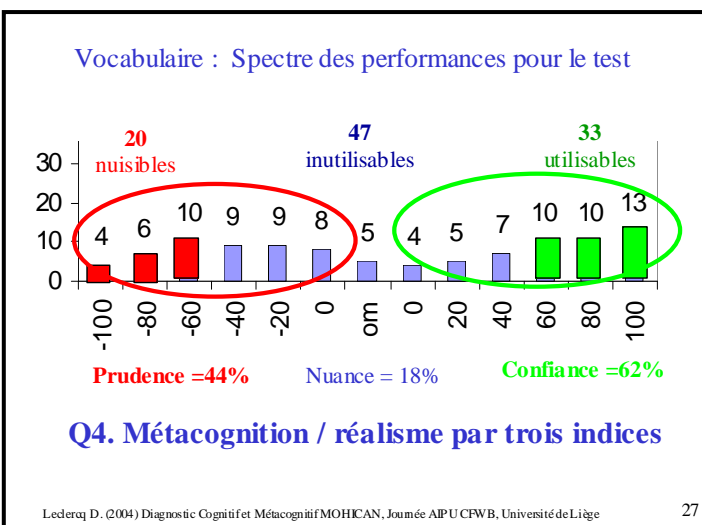
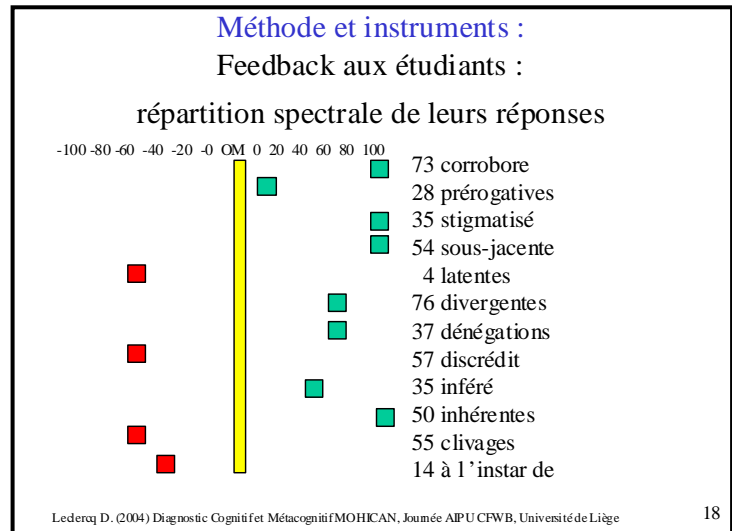
### B2.a) Evaluation à dimensions MULTIPLES. Exemple de l'exactitude + la concision

Lors de la résolution d'équations, Réginald Burton (1999) tient compte de l'exactitude de la réponse finale et du **nombre de transformations** effectuées pour y arriver. Il calcule un **score d'efficacité** grâce au nombre minimal d'étapes par lequel un expert peut résoudre le problème.

### B2.b) Evaluation à dimensions MULTIPLES. Couple 1 : Exactitude + Certitude

De Finetti (1965), Shuford (1966, 1999), Leclercq (1983, 1993, 2003), tiennent compte non seulement de l'exactitude des réponses, mais aussi des **certitudes** dans ces réponses, ce qui leur permet de calculer des **indices de réalisme** dans l'auto-estimation (sous-estimation ? surestimation ?), de dresser le « **spectre des performances** » combinant les 2 dimensions et de calculer un score composite « avec certitude ».

Dans l'opération MOHICAN (Leclercq, 2003) ou Monitoring Historique des CANDidatures, 4000 étudiants entrant dans 8 universités de la CFWB ont passé 10 tests, dont un sur le vocabulaire de la langue française (45 QCM à 5 solutions + Aucune et toutes). Chaque étudiant a reçu la position spectrale (de -100 à +100) de chacune de ses réponses (ci-contre pour 12 des 45 mots). La colonne (73, 28, 35, etc.) précédant le mot lui-même indique le % de Réponses Correctes dans la section de cet étudiant.

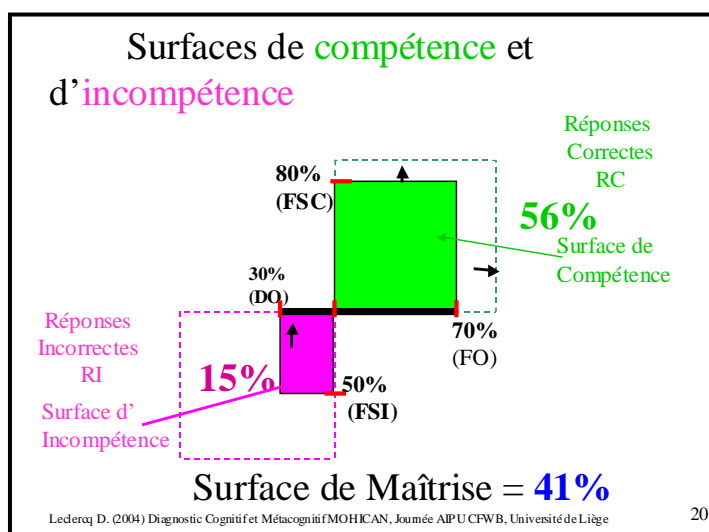
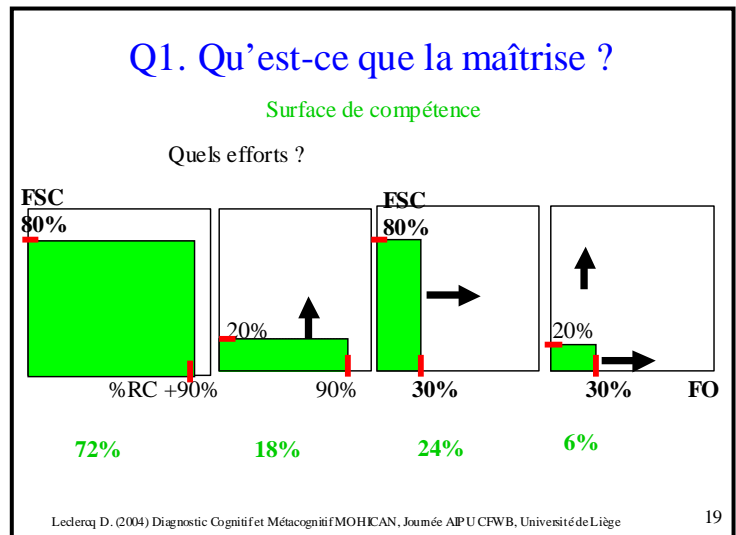


Une vue d'ensemble de ces positions (ou qualités) spectrales est aussi fournie. La courbe de l'hémispectre de gauche a une forme gaussienne ; celle de l'hémispectre de droite a une forme en J. Les indices de Confiance (certitude moyenne avec les réponses correctes), d'Imprudence-Prudence (certitude moyenne avec les réponses incorrectes) et de Nuance (différence entre Confiance et Imprudence) sont aussi fournis. (Leclercq & Poumay, 2005).

Habituellement, on exprime la compétence d'une personne dans une matière par son pourcentage de réussite à une épreuve dans ce domaine, le maximum possible étant 100%. Nous pensons qu'il faut tenir compte aussi de la certitude moyenne (voir Leclercq, 2003) des réponses correctes (ou confiance).

On calcule alors une surface de compétence, elle aussi exprimée en % (d'une surface idéale ou maximale).

Ainsi, un taux de Réponses Correctes de 90% accompagné d'une Confiance de 80% donne une surface de Compétence de 72% pour cette personne.



Quand il y a des réponses Incorrectes, il y a de l'incompétence.

Le taux de réponses incorrectes (ex : 30%) multiplié par l'Imprudence (ex : 50%) donne une surface d'incompétence de 15%, le maximum étant 100%.

La surface de maîtrise combine les deux :

$$\text{Surface de maîtrise} = \text{Surface de compétence} - \text{Surface d'incompétence}$$

Ici, **41%** = **56%** - **15%**

Ces concepts considèrent la compétence comme une variable bi-dimensionnelle (l'exactitude et la certitude) qui peut se représenter

- soit par un histogramme sous forme de distribution spectrale (axe horizontal de -100 à +100°
- soit par deux surfaces s'inscrivant comme deux parts de deux surfaces théoriques de 100%, l'une positive (la compétence maximale), l'autre négative (l'incompétence maximale).

### B2.c) Evaluation à dimensions MULTIPLES. Couple 2 : Exactitude + Rapidité

Dans certains domaines, le **temps de réponse** très lié à la qualité de la performance. Dans ces cas, le délai de réponse ou **indice de rapidité** est considéré comme un critère de qualité. Le temps de réponse (comme la certitude) porte autant sur le résultat que sur le processus.

Exemple 1 : Un test en temps limité.

Exemple 2 : Avec Merciadri (1982), nous avons testé des jeunes accouchées et mesuré le temps qu'elles mettaient à trouver le numéro du Centre Anti Poison dans l'annuaire téléphonique.

Exemple 3 : Avec Leblanc (1990), nous avons mesuré le délai de compréhension de dessins humoristiques sans parole (on reconnaît que la personne a compris le « gag » à son expression faciale.).

Exemple 4 : La Réanimation Cardio Pulmonaire (parfois appelée « bouche à bouche » est une opération qui, pour ramener une personne morte à la vie, doit réussir dans les 4 minutes !! (Sinon c'est inutile !)

Exemple 5 : Mathhues (1990) a mesuré le temps mis par une infirmière pour réparer la panne d'un a. un appareil palliant la détresse respiratoire (espère de poumon artificiel). Il est évident que, pour les mêmes raisons que dans l'exemple précédent, la vitesse d'intervention est cruciale.

Exemple 6 : En mathématique où Burton (1999) a observé que les réponses correctes sont fournies plus rapidement que les incorrectes. Fischer (1996, 83) a aussi observé que la détection d'expressions correctes (ex :  $7 \times 6 = 42$ ) elle aussi est plus rapide que celle des incorrectes (ex :  $8 \times 6 = 54$ ).

Exemple 7 : Vitesse d'amputation. Il y a plusieurs raisons de réduire la durée de l'amputation d'un membre : la réduction de la perte de sang, de la douleur pendant, de la durée des douleurs fantômes post-opératoires<sup>1</sup>, de l'indisponibilité du chirurgien (qui doit amputer d'autres blessés).

**Larrey** (Dominique-Jean, Baron (1766-1842), médecin de Napoléon.



*« suit les armées lors des guerres de la révolution et de l'Empire, en commençant par l'armée des Pyrénées, contre les Espagnols. A la bataille de la Sierra negra, il ampute ainsi en une journée 700 blessés. Plus tard, il met en place un système d'ambulances volantes »*  
(Récupéré de [http://fr.wikipedia.org/wiki/Dominique\\_Larrey](http://fr.wikipedia.org/wiki/Dominique_Larrey))  
(août 2005)

*« On l'a accusé d'avoir abusé des amputations. L'avenir lui donna raison car l'amputation précoce permettait de sauver près de trois-quarts des blessés et évitait la propagation du tétanos. Larrey écrit " Ce n'est pas au chirurgien de déterminer si une blessure est volontaire. Ce rôle appartient à un juge. Le médecin doit être l'ami de son patient. Il doit soigner le coupable aussi bien que l'innocent et concentrer ses efforts seulement sur la blessure. Le reste n'est pas son affaire."*

*À la bataille de Waterloo, Larrey fut pris par les Prussiens qui voulaient l'exécuter. Sa vie fut épargnée lorsqu'il fut reconnu par le Maréchal Blücher qui se souvenait que Larrey avait sauvé la vie de son fils en le soignant sur un champ de bataille quelques années plus*

*tôt. Ceci souligne une autre qualité de Larrey. Sur un champ de bataille, il soignait tous les blessés, qu'ils soient amis ou ennemis.*

*Un autre incident fameux se situe en 1830, durant la Révolution qui mit fin au règne de Charles X. Larrey refusa de livrer des Suisses à la vindicte des insurgés. Il alla au devant d'eux et cria: " Que voulez-vous ? Mes blessés? Ce sont les miens, laissez-les tranquilles". L'ayant reconnu, les rebelles pour toute réponse, l'applaudirent avec le plus grand enthousiasme.*

*Nombreux sont les historiens qui pensent que Larrey a été le moteur qui a entraîné la création de la Croix Rouge Internationale (1864) et la convention de Genève (1949) qui prescrit, entre autres, qu'un soldat ennemi blessé doit être soigné et qu'il est obligatoire de protéger les civils dans un territoire occupé. »*

**Algazi** in [http://www.napoleonicsociety.com/french/Larrey\\_Jean.htm](http://www.napoleonicsociety.com/french/Larrey_Jean.htm) aout 2005

<sup>1</sup> J-PH. Assal, Directeur à Genève du Centre de référence d'Education du Patient Diabétique de l'OMS, nous signalait que ces douleurs duraient pendant plus d'années que le temps d'amputation avait été long.

**Exemple 8 :**

**J.P. Fischer** (1996, 81-97) a développé, pour les 4 opérations arithmétiques (+, -, x, /) à l'école primaire une méthode appelée Juste-Faux qui recourt à un logiciel de sa conception. Ce logiciel affiche sur l'écran une opération complète (ex :  $9+7 = 15$ ) et l'élève doit taper sur la touche J (Juste) ou la touche F (Faux, ce qui est la Réponse correcte dans notre exemple F) dans un délai maximal de 5 secondes (le délai réel de réponse est enregistré). Les touches ont été recouvertes d'une gommette de couleur (verte pour J et rouge pour F pour minimiser l'interférence que constituerait la recherche des touches sur le clavier). Ce genre de test a été appliqué (Fischer, 2005) à 3 moments de l'année (mi septembre, fin janvier et mi juin) à 25 élèves d'une classe de CM2 (5° primaire en Belgique). Chaque test comporte deux séries, celle de niveau 1 avec des petits nombres (ex :  $2 \times 3 = 6$ ) et celle de niveau 2 avec des plus (ex :  $6 \times 8 = 48$ ).

Chacune des séries comporte 14 questions sur les additions (A), 14 sur les soustractions (S), 14 sur les multiplications (M), 14 sur les divisions (D). Pour chaque opération (A, S, M, D), on calcule le Temps de Réussite (TRu) sur les 12 meilleures performances (si 3 échecs ou plus, le score est blanc) et, suivant le résultat, on lui attribue un niveau de gris :

	3 échecs ou plus
	Tru entre 2,0 et 5,0 secondes
	Tru entre 1,7 et 2,0 secondes
	Tru entre 1,3 et 1,7 secondes
	Tru inférieur à 1,3 secondes

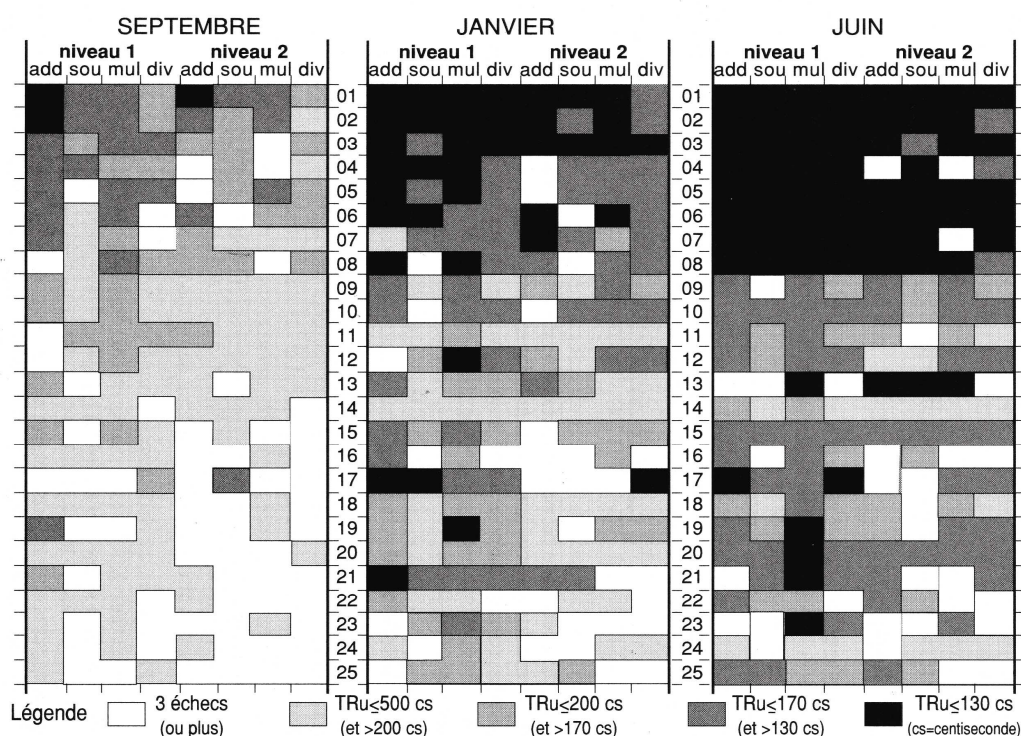
*Pour les différences entre élèves, les images sont d'une telle netteté qu'elles n'ont pas besoin d'être commentées. Pour les différences entre opérations, on peut par exemple voir que les soustractions de niveau 1 sont beaucoup moins bien maîtrisées que les multiplications de même niveau. » (p. 87)*

*« La visualisation donne une impression de progrès général et important, dans le sens où il concerne tous les élèves et toutes les opérations et niveaux. Elle suggère clairement qu'aucun élève performant en début d'année (haut de l'image de septembre) n'avait atteint un « plafond » l'empêchant de progresser par la suite....*

*On peut voir que certains élèves ont –les numéros 15 et 20 notamment- ont quand même plus progressé que d'autres –les numéros 14 et 24 par exemple.*

*Enfin, ... on peut entrevoir un ralentissement du progrès dans la deuxième moitié de l'année, eut-être en partie dû à une démobilitation de certains élèves à la veille des grandes vacances.*

*Quel est l'origine de ce progrès massif ? On peut, là aussi, penser à un « effet grandes vacances » (88-89).*



### Exemple 9 : vitesse de lecture

**Jean Foucambert** est l'inspirateur des logiciels (et concepts) ELMO (Entraînement à la Lecture par micro-ordinateur) et ELSA (Entraînement à la Lecture Savante). ELMO présente une grande variété d'exercices sur de très nombreux textes. Par exemple, des exercices ayant pour but d'augmenter **l'empan de lecture** (le nombre de mots ou de lettres saisis dans une seule fixation oculaire), très lié à la vitesse de lecture. Voici quelques-uns des repères qu'il fournit dans **La manière d'être lecteur** (Paris : Albin Michel) :

« Comment interpréter les résultats en **vitesse** :

1- En dessous de 550 signes/mn, le lecteur est au stade de l'exploration syllabique, et il a beaucoup de difficulté à trouver du sens.

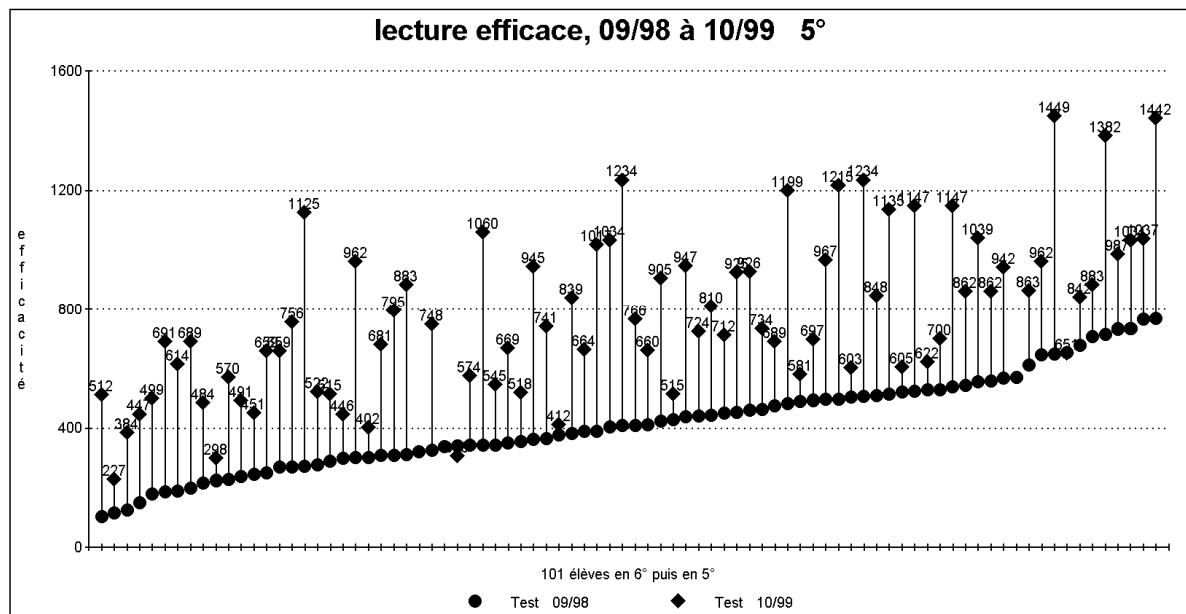
2- Entre 550 et 820 signes/mn, le lecteur arrive difficilement à construire un sens. Les meilleurs de cette catégorie sont de bons déchiffreurs, mais pas encore des lecteurs.

3- Entre 820 et 1300 signes/mn, on peut véritablement parler de lecture. La compréhension s'effectue sans problème.

4- Au-delà de 1300 signes/mn, la lecture est bien installée.

Les élèves des deux premières catégories ont une vitesse inférieure à celle requise pour accéder à une lecture digne de ce nom. Ce sont des "liseurs" et ils ont toutes les chances de le rester si une action spécifique n'est pas mise en place pour les faire progresser. »

In <http://bigonnet.club.fr/Foucambert.html> (août 2005).



Ce graphique montre les résultats de deux tests de lecture efficace, pour un groupe de 101 élèves de 6ème que je suis en 5. Le test de septembre 98 est symbolisé par un rond noir, celui d'octobre 99 par un losange.

Parmi les méthodes de mesure de la vitesse de lecture, citons les travaux de Jourdain, Zagar et Lété (2005).



**Jérôme Kagan** (1965) s'est intéressé au « tempo cognitif » c'est-à-dire à l'opposition impulsivité / réflexivité (les termes sont de lui).

Les réflexifs réussissent mieux que la moyenne mais dans un temps supérieur à la moyenne. Les impulsifs réussissent moins bien que la moyenne et dans un temps inférieur. Les deux autres groupes, sans nom, pourraient être appelés « les meilleurs » et « les pires ».

Temps de réponse

		< M	> M
%	< M	impulsifs	pires
RC	> M	meilleurs	réflexifs

Son *Familiar Figures Matching Test* (Test d'appariement de figures familières) a été conçu pour classer les personnes dans ces 4 catégories. En voici une question :

### B.2d) Les examens en temps limité

#### a) Temps d'ensemble limité

Nombre d'épreuves écrites se déroulent en temps limité (ex : un test contenant 60 QCM à passer en 2 heures), à charge pour l'évalué de répartir ses efforts de façon efficace. Les étudiants habiles à passer des tests (qui font preuve de « *test wiseness* ») analysent rapidement chaque question et répondant d'abord celles pour lesquelles ils connaissent bien la réponse, laissant les autres pour la suite, plutôt que de se bloquer sur une seule question, ce qui risquerait de consommer tout leur temps.

#### b) Temps limité par parties

Cet aspect du test wiseness ne se pose pas dans les évaluations de type ECOS (Evaluations Cliniques Objectives et Structurées), particulièrement développées en médecine (Bourguignon, 1997) où l'étudiant devra passer dans diverses « stations ». Chacune, avec son examinateur, sert à évaluer une performance simulée précise. Par exemple : anamnèse d'un patient, diagnostic après entretien fictif par téléphone, examen corporel d'un patient, injection, bandage, etc.

La durée de l'épreuve dans chaque station est en temps limité: toutes les 6 minutes. Quand retentit la sonnerie, chacun des étudiants doit sortir et se rendre dans la station suivante (rotation), et ainsi de suite, jusqu'à la dernière.

### B3. Un exemple d'évaluation de la vitesse et de la certitude.

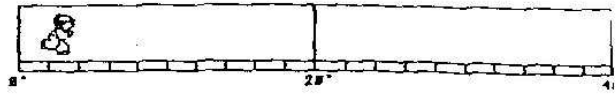
Un premier exemple concerne l'évaluation des compétences techniques d'enfants et d'adolescents diabétiques (Ackermans, Ernould, Leclercq, 1985). L'apprenant a le choix du domaine (injections, physiologie, activités physiques, analyse d'urines, ...) dans lequel il se fait poser des questions (puisées au hasard dans une "banque").

A chaque question, le didacticiel (MASTER DIAB) exige de l'enfant qu'il réponde en moins de 40 secondes (avec un "bonus" s'il répond en moins de 20 secondes).

En outre, l'enfant doit exprimer son degré de certitude.

N.B. : Les couleurs sont celles que prend le CLINITEST trempé dans l'urine (bleu = parfait; rouge = alarme).

La fonctionnalité de ces deux exigences a été dictée par les diabétologues eux-mêmes. On sait en effet que la réaction de la personne doit survenir avant la perte de conscience (coma hypoglycémique par exemple) et que la certitude quant à la nature de l'action à mener (injecter ou non ? quelle insuline ? quelle quantité ? à quel endroit du corps ? etc.) importe.



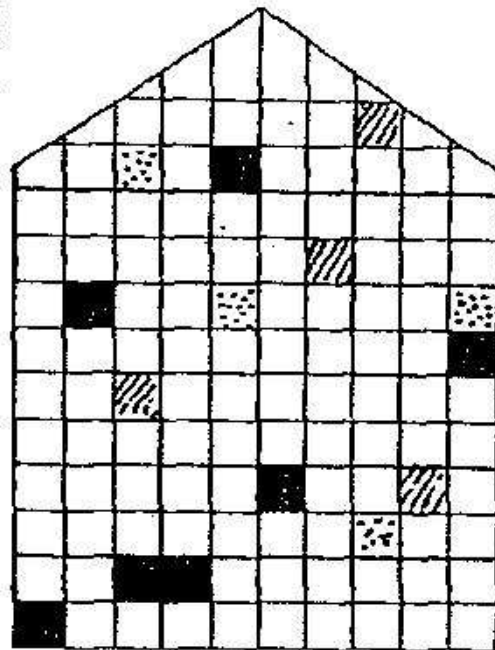
Tu disposeras de 40 secondes pour répondre. Rappelle-toi la séquence précédente. Pour te les rappeler, je vais courir sur la piste ci-dessus. Après 20 secondes tu entendras un bip. Le temps écoulé, tu entendras 2 bips.



Tu as sûrement remarqué que les briques sont de 4 couleurs différentes.

- les bleues représentent les questions parfaitement réussies.
- les vertes représentent les questions bien réussies.
- les jaunes représentent les questions réussies mais peu sûres.
- les rouges représentent les questions ratées.

L'idéal est d'avoir une maison toute bleue.



Appuie sur NEXT



## C. Les SUJETS de l'évaluation : QUI ?

### C1. Evaluation portant sur des PERSONNES vs sur des INSTITUTIONS

#### a) L'évaluation de personnes (seules ou en groupe)

Les tests et examens portent souvent sur les performances et compétences de personnes ou de groupes d'individus, selon des critères de présages, de processus ou de produits (cf. ci-après).

#### b) L'évaluation institutionnelle

« *Evaluation d'une institution d'enseignement réalisée le plus souvent en fonction de critères spécifiques : conditions pour l'octroi de subventions, pour être reconnu comme établissement pilote, etc.* » (de Landsheere, 1979, p.113)

On peut trouver des critères de

Présage : la taille de l'établissement, le nombre d'enseignants détenteurs d'un diplôme d'un certain niveau, le nombre d'heures de cours données, les infrastructures (restaurant, gymnase, douches, etc.), les ressources techniques (laboratoires, nombres d'ordinateurs),

Processus : modalités de fonctionnement de la gestion, relations avec les délégués syndicaux et l'association de parents, taux d'absentéisme des étudiants, des professeurs, etc.

Produits : taux de réussite à la sortie (ratios entrées / sorties), taux de réussite ou d'emploi ultérieur de ceux qui sont sortis,

L'approche « Qualité » amène à l'évaluation des universités, selon une démarche transactionnelle où une auto-évaluation précède un audit externe (Romainville et Boxus, 1998,13-32).

On utilise de plus en plus le principe d'une « Analyse SWOT » (Strengths, Weaknesses, Opportunities, Threads). En voici une ne portant que sur une opération institutionnelle.

**RESSAC** : a strategy to improve learning strategies to achieve learning outcomes  
Faculty of Psychology and Education of the University of Liège  
D. Leclercq, S. Bredart, M. Crahay & Ch. Mormont

In numerous faculties, university professors are deceived about the high rate of first year students who **study in an inadequate way**, not realizing that they are **expected to know AND to understand**. Some know but do not understand since they have studied in a superficial way and stored contents by rote learning<sup>1</sup>. Some others understand, practiced deep learning, but did not make the effort to memorize since they did not realize that in each science there are fundamental facts or methods that have to be mastered. Not mentioning those who did half of each.

Professors know that for a long time and the institution has developed actions to prevent it, by offering lectures, seminars and exercises on how to study, assuming principally that there is **a lack in their capacity to memorize and understand**. Results are deceptively low in terms of changes.

An other way to face the problem is to make the hypothesis that for a lot of students **the lack is** not in their capacity but **in their will to memorize, or to deep learn or both**, that many of them try how their spontaneous way to study will work (pass or fail) and will change only if they fail. Starting from this standpoint, an experiment has been organized for the first year students of the faculty of psychology and Education of the University of Liège to take advantage of the first real size feedback, i.e. just after the mid-term exams where students usually are informed of they successes and failures in 4 or 5 courses partial exams. Since they will have a second chance (in June or September), the feedback is of great importance for those who failed (the majority of them). Classically, the feedback is a score on 20 for each of the courses. With such a minimal feedback, students authorize themselves

<sup>1</sup> (Entwistle, Houssell & Marton, 1984)

to **attribute their failures to external causes**<sup>2</sup> such as excessive severity of a teacher, excessive difficulty of a content, excessive complexity of questions, momentary disease when taking the test, etc.

In 2000 for the first time and this year (2005) for the second time, 4 professors decided to split their score (on 20) on **two distinct scores**, one related to **memory** (Knowledge score) and one to **understanding** (Use of Knowledge score). These feedbacks were also given in a graphical way, called the Z radiography, where the 8 scores were displayed in horizontal histograms centered on 0, the average score and ranking from -3 to +3, i.e. in Z scores.

Therefore, each student could see, from this **personal radiography** whether there were cross courses tendencies such as “below average in 3 courses out of 4 for understanding, but above average in memory”.

These pieces of information were delivered before the month dedicated to prepare the exams of June via autonomous study (courses are finished). Interviewed after the June exams but before knowing their June results, students had to tell **whether they changed their study methods** (and in which respect). Some did, others did not. These two a posteriori groups were compared in terms of successes and failures in their June and September exams. **The difference in terms of successes** was dramatically in favour of the group of students who changed their study method accordingly to their Z radiography, in each of the 4 courses.

The experiment and the results have been largely described<sup>3</sup>.

#### RESSAC

Strengths	Weaknesses
<ul style="list-style-type: none"> <li>-Does not require excessive extra work from teachers if scores are obtained by automatically scored exams (MCQs, etc.)</li> <li>-The feedback is simple since limited to two concepts (memorization – understanding)</li> <li>-The feedback is repeated (on 4 courses) so that the student can make the difference between systematic features and occasional ones.</li> <li>-The feedback is directly linked to the study method that is in the students’ hands. It empowers them by understanding the effects of their decisions (it makes the causes internal and changeable).</li> </ul>	<ul style="list-style-type: none"> <li>-The strategy must have been planned in advance in order to insure a minimal number of questions revealing memorization or deep understanding, since some questions are so “undecidable” that they are not taken into account to compute any of the two subscores. –This requests additional work when the copies are open ended essay type questions.</li> <li>-A part of the effectiveness of this approach is due to the fact that the exams count: they REALLY fail or pass. Earlier in the year, knowing that the test is just formative, the students would care less.</li> </ul>
Opportunities	Threads
<ul style="list-style-type: none"> <li>-The issue is shared by many faculties</li> <li>-The procedure can be applied in any faculty</li> <li>-Optical reading systems and On-line formative testing can offer this possibility earlier in the school year.</li> </ul>	<ul style="list-style-type: none"> <li>-Students could live this approach as “invasive” (whereas we did not observe a single complain of this kind)</li> </ul>

<sup>2</sup> (Rotter, 1966 & Wiener, 1985)

<sup>3</sup> Leclercq (2003). RESSAC : Résultats d’Epreuves Standardisées au Service des Apprentissages en Candi<sup>3</sup>. In Leclercq (Ed). Diagnostic cognitif et métacognitif au seuil de l’université. Liège : Editions de l’Université de Liège, 155-170.

## C2. Evaluation INDIVIDUELLE vs DE GROUPE

### a) L'évaluation individuelle

D'habitude, l'évaluation porte sur les performances d'UNE personne, au point que des dispositions sont prises pour interdire, dissuader et empêcher la collaboration. Par exemple, au Centre d'Auto Formation et d'Évaluations Interactives Multimédias (CAFEIM) de la faculté de Psychologie et Education d l'Université de Liège, des « flasques » sont sorties des tables pour servir de séparation entre étudiants lors d'examens sur ordinateur.



### b) L'évaluation d'un groupe

Il ne manque pas, dans la vie, de situation où c'est la performance indissociable du groupe qui est évaluée. C'est le cas pour une course d'aviron en huit avec barreur, pour les courses par équipe en cyclisme (où c'est le temps du dernier qui est pris en compte), pour le double en tennis, etc.

Il arrive que des étudiants se plaignent d'être « mis dans le même sac » que leur(s) partenaire(s) sur la base de l'évaluation d'un travail collectif commun. Diverses parades permettent de prévenir cette objection fondée:

Un commentaire « personnel » est demandé à chacun des co-auteurs sur le processus et/ou le résultat du travail, sur base, éventuellement, de son journal de bord personnel (EDUM & DOCIMO chap.3, K3, p.29).

L'évaluation des personnes sur base d'un produit de groupe.

Le professeur peut réunir les auteurs et leur poser, tout à tour, les questions sur le travail commun écrit.

Dans cette formule, chaque auteur du rapport est « réputé » avoir tout réalisé seul même si, dans la réalité, l'exécution des tâches ont été partagées ; autrement dit la responsabilité de la conception repose également sur les épaules de chacun.

On peut aussi imposer l'obligation de partager les acquis avec les autres membres du groupe.

En 2003, nous avons organisé des PARMs collectifs où chaque équipe comportait 5 étudiants. Ils avaient à respecter 10 critères (voir colonne 1 des tableaux), mais pouvaient s'organiser à volonté pour se répartir le travail. Cette répartition (tableau 1) est indépendante de la notation du professeur. , Le professeur cote non pas chaque étudiant (il ignore qui a fait quoi), mais chaque critère. Ensuite seulement, le professeur donne son évaluation et les étudiants leur répartition (photo).



Les étudiants notaient leur répartition des tâches dans leur colonne (E1 à E5), chaque colonne devant totaliser 100% et chaque ligne 50% (puisque les 100% des 5 étudiants devaient être répartis sur 10 critères).

Tableau 1

Groupe X	Notes	sur	E1	E2	E3	E4	E5	Max Ligne 50%
			Max Col. =100	Max Col. =100	Max Col. =100	Max Col. =100	Max Col. =100	
Défi		10	10	10	10	10	10	50
Exposé Fond		10	10	10	10	10	10	50
Exposé Forme orale		10	25				25	50
Exposé Forme Média		10			25	25		50
Critique		10		50				50
Lien		10			25	25		50
Activité-fond		10	10	10	10	10	10	50
Activité forme		10	25				25	50
Réponse aux Q fond		10	10	10	10	10	10	50
Réponse aux Q forme		10	10	10	10	10	10	50
Somme		100	100	100	100	100	100	500
				Leclercq,	Micro PARMs	MFPA, STE	ULG, 2002	35

Les notes du professeur combinées à la répartition des tâches entre étudiants peuvent déboucher sur des notes finales différentes pour différents étudiants (tableau 2).

Scores pondérés														
Titre			Etudiant 1		Etudiant 2		Etudiant 3		Etudiant 4		Etudiant 5		Total Implic max 50	
Groupe	Date	sur	Implication	Points	Implication	Points	Implication	Points	Implication	Points	Implication	Points		
Définition du défi	Fond	4	10	10	40	10	40	10	40	10	40	10	40	50
Exposé	Fond	8	10	10	80	10	80	10	80	10	80	10	80	50
	Forme orale	6	10	25	150	0	0	0	0	0	0	25	150	50
	Forme média	7	10	0	0	0	0	25	175	25	175	0	0	50
Critique -Lien	Fond	7	10	0	0	50	350	0	0	0	0	0	0	50
	Forme	6	10	0	0	0	0	25	150	25	150	0	0	50
Activités	Fond	8	10	10	80	10	80	10	80	10	80	10	80	50
	Forme	7	10	25	175	0	0	0	0	0	0	25	175	50
Réponses aux Q des étudiants	Fond	8	10	10	80	10	80	10	80	10	80	10	80	50
	Forme	8	10	10	80	10	80	10	80	10	80	10	80	50
Total		0	100	100	685	100	710	100	685	100	685	100	685	
					13,7		14,2		13,7		13,7		13,7	

Groupe N Scores pondérés

Leclercq, Micro PARMs MFPA, STE ULG, 2002

36

## D. LES DESTINATAIRES de l'évaluation : POUR QUI ?

### D1. Evaluation à destination CONFIDENTIELLE-PRIVEE vs PUBLIQUE

#### a) Résultats à destination privée-confidentielle

L'évaluation formative est le plus souvent privée ; {c'est une} *sorte de dialogue particulier entre l'éducateur et son élève* (de Landsheere, 1979, p. 115).

Les résultats (question par question) au Check-up MOHICAN (Monitoring Historique des CANDidatures), ont été communiqués confidentiellement sous pli fermé à chaque étudiant, ceux de chaque faculté à chaque Doyen, ceux de chaque université à chaque Recteur. (Leclercq, 2003).

#### b) Résultats à destination publique

La communication des notes a parfois la publicité des bans de mariage. Ainsi à l'université de Durham les notes sont exposées sur la place publique.

#### c) Situations intermédiaires (à confidentialité restreinte)

Le bulletin scolaire traditionnel est semi public car connu au moins des parents de l'élève (il est d'ailleurs fait pour cela).

Le score personnel au TOEFL (*Test Of English as a Foreign Language*) est communiqué confidentiellement à l'intéressé (qui pourra en faire état où il le juge utile), mais est demandé à l'entrée de certaines universités américaines, ce qui le rend dès lors semi-public (si non connu des autres candidats) quant aux destinataires. Sont aussi semi publics les résultats aux *Medical Boards Examinations* que passent les médecins désireux de pratiquer dans un autre état (des USA) que celui de leur diplôme.

#### d) Exemple de destinataires multiples

A l'Université de Liège, les enseignements sont évalués par les étudiants au moyen de questionnaires «EVALENS» (Gilles, 1999). Les différents feedbacks ont des degrés de confidentialité différents.

Les étudiants voient, pour chaque question (ex : « L'exposé (oral) est bien structuré »), la distribution des scores moyens des 8 enseignants (allant de « -3 Pas du tout » à « +3 Tout à fait »).

L'exposé est bien structuré	-3	-2	-1	+1	+2	+3
		2	1	1	2	2

Les étudiants peuvent voir que 2 seulement (des 8) sont considérés comme « au sommet » de la qualité, aucun n'étant classé au minimum.

Chaque professeur reçoit la même distribution, mais peut s'y situer car sa position est grisée :

L'exposé est bien structuré	-3	-2	-1	+1	+2	+3
		2	1	1	2	2

Dans l'exemple ci-dessus, le professeur peut constater que le jugement est « plutôt d'accord », qu'un autre collègue a reçu le même jugement moyen de la part des étudiants, et que deux collègues ont reçu une meilleure appréciation que lui (elle), mais il (elle) ne sait pas qui sont tous ces collègues.

Le Doyen de la faculté voit, lui, la position de chaque enseignant avec son nom dans la distribution.

L'exposé est bien structuré	-3	-2	-1	+1	+2	+3
Professeur AMPERE		2	1	1	2	2
Professeur BRASSENS		2	1	1	2	2
Professeur COLUCHE		2	1	1	2	2
Professeur DESCARTES		2	1	1	2	2
Professeur EINSTEIN		2	1	1	2	2
Professeur FERMI		2	1	1	2	2
Professeur GARIBALDI		2	1	1	2	2
Professeur HERMES		2	1	1	2	2

Le Doyen félicitera sans doute BRASSENS et EINSTEIN et demandera un effort à COLUCHE et HERMES.

## E. Les AGENTS de l'évaluation : PAR QUI ?

### E1. Evaluation à cible

#### ALLO- vs AUTO-CENTREE

On se focalise ici sur le QUI (est évalué) et PAR QUI : l'évaluateur et l'évalué sont-ils une seule et **même** (αυτο en grec) personne ou une **autre** (αλλο en grec) ?

#### a) Evaluation allo-centrée (d'autrui)

C'est le cas le plus banal du formateur évaluant le formé, de l'expert évaluant le novice, du sélectionneur évaluant le candidat, etc.

#### b) Evaluation auto-centrée (ipsative)

C'est le cas de l'apprenant qui relit SA copie, ou qui estime (sur une échelle de certitudes) son degré de confiance dans SA réponse (qu'il vient de donner), par exemple le nombre de fautes laissées dans la dictée qu'il vient de terminer. Il s'agit là de **post-diction** de son résultat (car estimation après la réponse).

Quand il juge de SA capacité à répondre dans un domaine en général (avant de connaître les questions), il s'agit de **prédiction**.

N.B. : Auto-évaluation peut s'interpréter de multiples façons :

Evaluation de ou sur soi-même ou encore évaluation **ipsative** (ex : estimation de ses propres capacités, donc **auto-estimation**).

Evaluation de soi-même sans aide (de façon autonome); il serait alors moins ambigu de parler d'**autorégulation** ou de régulation autonome ou indépendante (voir l'Introduction du présent document). C'est ici le côté autonomie qui est souligné.

Selon Partlett, l'évaluation illuminante porte sur ce que les sujets (les acteurs) ressentent (de Landsheere, 1979, 114). A eux donc de l'exprimer.

### E2. Evaluation à exécution

#### AUTONOME vs HETERONOME

En Grec ancien, nomè signifie la loi, la règle. Autonome signifie « qui a fixé la règle lui-même ». Hétéronome signifie : « qui applique la règle fixée par autrui. »

#### a) L'évaluation hétéronome

Dans la régulation (voir Introduction), les opérations

- (1) de définition des objectifs à évaluer,
  - (2) de choix des questions,
  - (3) de choix du moment,
  - (4) de définition des critères,
  - (5) de correction des réponses,
  - (6) de décisions en conséquence
- (Etc.) sont menées

- soit par l'étudiant lui-même (auto-servuction)
- soit par une autre personne (allo-servuction).

L'hétéronomie et l'autonomie portent sur un sous-ensemble de cela, et concerne les seules opérations où une décision est à prendre, c.-à-d. quasi toutes celles ci-dessus (sauf la 5..et encore).

La mise d'une banque de questions à la disposition des étudiants (Gilles, 1998) leur permet, au moins, de gérer le moment et (pas toujours) la durée ou (pas toujours) l'ordre-de-passation de l'épreuve.

#### b) L'évaluation autonome

Elle suppose que des conditions de travail autonome soient réunies :

- visibilité des objectifs
- existence d'épreuves auto-administrables
- critères de correction auto-applicables
- consignes d'interprétation des résultats
- barème d'excellence
- propositions de remédiation

Bref des « Modules d'Auto-Evaluation » (Leclercq, 1985).

**c) Les évaluations partiellement autonomes** sont celles où l'étudiant applique des critères d'évaluation qu'il a conçus AVEC le professeur.

### E3. L'évaluation par les EXPERTS ou par les PAIRS

#### a) L'évaluation interne par EXPERTS

C'est celle que pratiquent les enseignants qui jugent les étudiants qu'ils ont formé eux-mêmes. On pourrait aussi parler d'évaluation **participante**, comme l'anthropologue Malinowski parlait d'observation participante qui « amène le chercheur à vivre la vie des groupes qu'il étudie, à partager le plus possible leurs activités pour mieux comprendre leur vision du monde (de Landsheere, 1979, p. 192).

#### b) Evaluation externe par EXPERTS

« Evaluation réalisée par des personnes ne faisant pas partie de l'équipe éducative chargée de réaliser un programme. » (de Landsheere, 1979, p.113)

Des examens externes caractéristiques sont

##### En France :

- le Baccalauréat,
- les concours d'entrée aux grandes écoles (Polytec, ENA, ENS, EMP, EMN<sup>1</sup>), etc.

##### Aux Etats Unis :

- les tests passés à la charnière secondaire/universitaire tels que le SAT (Scholastic Aptitude Tests),
- les Medical Boards (tests que passent les médecins pour être autorisés à exercer la médecine dans un autre Etat que celui où ils ont reçu leur diplôme)

##### En Belgique :

- les examens cantonaux en 6<sup>o</sup> primaire,
- les examens du Service de Sélection et Orientation (SELOR) de l'Etat,
- les examens d'entrée aux études d'ingénieur.

#### c) Situations intermédiaires

A la Faculté de médecine de l'Université de Maastricht, les professeurs affectés à la construction des Progress Tests n'enseignent pas et ceux qui enseignent ne testent pas.

#### d) Evaluations par les PAIRS

Ce processus où des groupes d'individus apprécient leurs pairs, est surtout efficace lorsque les critères d'évaluation sont bien explicités (Falchikov, 1995, 175). Les contextes de « supervision » (dans un stage) ou de « mentoring » (par un étudiant plus âgé) se prêtent particulièrement à ce genre d'évaluation.

Dans la plupart des expériences décrites dans la littérature, ce qui est ainsi évalué ce sont les performances (médicales, pédagogiques, relationnelles) plutôt que les connaissances.

Voici, par exemple, les 4 grands critères d'évaluation par les pairs d'une présentation orale :

- Structure et cohérence ;
- Connaissance du contenu ;
- Quantité appropriée d'information ;
- Clarté et expressivité dans l'exposé.

Dès 1971, Korman & Stubblefield observaient que l'évaluation (*rating*) par les pairs constituait le meilleur prédicteur de la performance (médicale) d'interne, et ce probablement parce que les étudiants se connaissent les uns les autres sous un angle difficilement accessible par le professeur (Linn *et al.*, 1975).

D'Augelli (1973) constate que l'évaluation par les pairs est bien corrélée avec celles des experts en ce qui concerne les « comportements interpersonnels », ce que n'est pas l'auto-évaluation. Boud & Tyree (1979) observèrent des corrélations pairs / experts très élevées (0,75 et 0,83), mais ipso / experts plus faibles.

Les bénéfices ressentis (déclarés par les étudiants objets de l'évaluation par leurs pairs) sont « une amélioration de ma performance d'examens en comprenant mieux ce que les examinateurs prennent en considération dans les réponses » (Magin & Churches, 1988) et « Cela me force à structurer plus mes réponses et à étudier plus » (Falchikov, 1986).

---

<sup>1</sup> Ecole Polytechnique (de l'armée), Ecole Normale d'Administration, Ecole Normale Supérieure, Ecole des Mines de Paris, Ecole des mines de Nancy.

## E4. Evaluation dans un rapport TRANSACTIONNEL vs ARBITRAIRE

### a) Evaluation arbitraire

L'évaluation « vient d'en haut » et il y a peu de place pour une « discussion entre l'évaluateur et l'évalué ».

### b) Evaluation transactionnelle

Une telle évaluation se produit chaque fois que les évalués peuvent faire entendre leur point de vue ...et que celui-ci est réellement écouté...avec effet possible sur la note. Dans le processus – très actuel – d'évaluation de la qualité dans les universités, on procède en deux temps. D'abord, l'université établit un rapport d'auto-évaluation qui est adressé à des experts extérieurs qui, sur cette base, procéderont à un audit sur place. On voit que c'est surtout à leur cohérence » entre LEURS objectifs, LEURS stratégies et LEURS résultats) que sont ainsi jugées les universités.

Dans une expérience plus modeste, Jans et Leclercq (1998) ont demandé à leurs étudiants d'auto-évaluer leur travail (Rapport écrit sur une expérience personnelle) à 8 points de vue. L'encadrant a ensuite évalué les mêmes travaux sur les mêmes critères, sans connaître les auto-évaluations. Un dialogue transactionnel s'est alors engagé, en se concentrant sur les notes divergentes entre l'auto et l'allo-évaluation.

Des révisions ont eu lieu dans les deux sens. (Leclercq et Poumay, 2004, chap. 7).

M. Serrurier (2005) a pratiqué de la sorte (photo ci-dessous).

### c) Situations intermédiaires

Législation belge prévoit

- le droit pour l'étudiant de voir sa copie
- l'exigence de recorection en cas d'erreur de jugement.



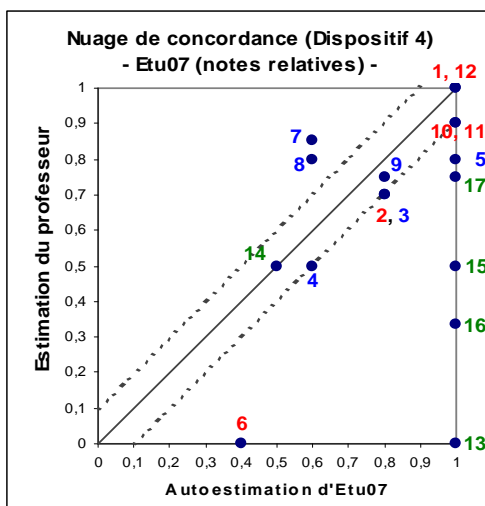


#### d) L'expérience de confrontation systématique de Jans (2000)

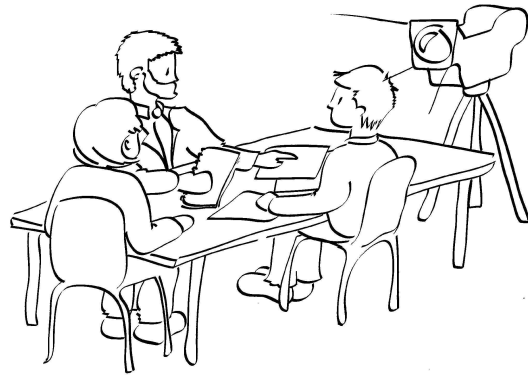
Elle portait sur la **confrontation dialectique des jugements** entre l'étudiant et le professeur sur des PARMs (voir Jans et al., in Leclercq, 1998, chap. 9). Les notes d'auto-évaluation ont été remises **sous pli scellé** afin que le professeur ne puisse pas en prendre connaissance avant de mettre ses propres notes sur le même travail. Le professeur a annoncé que l'entretien se déroulerait comme suit :

« Les 17 critères seront passés en revue, mais nous discuterons principalement ceux pour lesquels existe un écart important entre votre note et la mienne, et donc une discordance de vues entre vous et moi. »

Pour les performances complexes, voici comment Jans (2000) a présenté les auto-notations ou jugements (en abscisse) comparées aux allo notations ou jugements (en ordonnée), à gauche pour un étudiant sur les 17 critères (numérotés), et à droite pour les 18 étudiants pour un critère précis « Présentation de l'article ». On constate que l'étudiant 07 est plus optimiste que le professeur. Une confrontation professeur / étudiant a eu lieu, créant ainsi une situation POST. Sans une telle confrontation et discussion, les discordances sont considérées comme des erreurs d'auto-évaluation de l'étudiant.



Le graphique individuel (de gauche) de chaque étudiant lui a été remis pour qu'il prépare sa rencontre-discussion (confrontation dialectique) avec le professeur.



Lors de la séance de confrontation, l'étudiant et le professeur sont munis des notes (et justifications de ces notes) de l'un et de l'autre. « L'étudiant et le professeur ont eu l'occasion de se poser mutuellement des questions, de demander des explications quant aux points accordés ou aux justifications écrites fournies. » (p. 345). Ces entretiens ont été filmés.

Jans (2000, 336) rapporte la phrase de Boud (1995, 204-205) sur le « feedback réussi » :

« Si vous souhaitez donner un feedback efficace, vous devriez : être réaliste, être précis, être sensible aux objectifs de la personne, répondre en temps opportun, être descriptif, ne pas porter consciemment de jugements, ne pas comparer, vous montrer appliqué, être direct, être positif, être conscient de votre état d'esprit. ». Jans continue : « Et au récepteur du feedback, Boud lui conseille d'être explicite<sup>1</sup>, attentif, conscient de son état d'esprit<sup>2</sup>, silencieux. ».

<sup>1</sup> Make it clear what kind of feedback you are seeking

<sup>2</sup> Notice your own reactions, both intellectual and emotional

## F. Les tempos de l'évaluation : QUAND ?

### F1. Evaluation à périodicité CONTINUE vs PONCTUELLE

#### a) L'évaluation continue

« Collecte systématique de scores ou d'appréciations...aboutissant à une note finale. Peuvent intervenir dans l'élaboration de cette note : les travaux faits en classe, les travaux à domicile, les résultats d'interrogations écrites, de tests, ainsi que le jugement subjectif de l'enseignant. L'évaluation continue est...un processus cumulatif, suivant le développement de l'élève et réfléchissant les changements qui interviennent dans ses réactions au cours. » (de Landsheere, 1979, p. 112).

Contrairement à l'évaluation répétée (voir G2), ce ne sont pas les mêmes épreuves qui sont utilisées en évaluation continue (ou « diluée » ou encore « répartie ou fractionnée » dans le temps).

#### b) L'évaluation ponctuelle ou concentrée

« Evaluation effectuée à un moment donné pour répondre à une question à propos d'un individu ou d'un programme (passage d'une classe, attribution d'une bourse,...). Ce type d'évaluation ne semble se justifier pleinement que dans les cas où une évaluation continue est impossible. » (de Landsheere, 1979,114)

C'est le cas le plus souvent, du mémoire de fin d'études universitaires, car les étudiants n'ont jamais eu auparavant l'occasion de mener à bien un travail d'une telle ampleur et présentant de telles exigences et la plupart d'entre eux ne produiront pas de doctorat. On s'efforce, cependant, de les y préparer par des travaux similaires « en réduction ».

Il existe des situations où le formé lui-même souhaite être libéré d'une évaluation continue, par exemple dans l'enseignement supérieur. Il n'est donc pas rare de voir des institutions offrir le choix à leurs étudiants : SOIT une évaluation « étalée » (dans l'enseignement universitaire par des « partiels dispensatoires », SOIT une évaluation « concentrée » (examen durant la session).

### F2. Evaluation à occasion(s) ou épreuve(s) UNIQUE vs REPETEES

#### a) Evaluation unique

Il arrive qu'une performance soit évaluée une seule fois dans la vie d'une personne et à un moment donné (il n'y a ni repêchage, ni deuxième session). Cette évaluation est évidemment ponctuelle. C'est le cas de la défense de doctorat en Belgique (alors qu'aux USA, le processus est plus « continu ». En pédagogie, les « retests » sont difficiles, car le simple fait de passer le test produit un apprentissage.

Par retest, G. de Landsheere (1979, p. 238) entend « le test répété dans les mêmes conditions et à l'aide du même instrument ou d'une forme parallèle (équivalente) ».

#### b) Occasions ou épreuves répétées

On peut raisonnablement donner un même test

1. quand on pense que les délais (plusieurs mois) et le jeune âge font oublier les « réponses correctes ».

2. quand c'est ce qu'il faut maîtriser, quand c'est la population des questions (ex : l'alphabet, les temps primitifs, les tables de multiplication, les connaissances sur le diabète pour un patient) et qu'on peut se permettre (on a le temps) de poser toutes les questions.

Idem pour le bilan quotidien d'une journée de travail (ou de loisirs) en groupe (ex : lors du « Conseil du soir » à un camp scout), fait appel aux mêmes questions chaque soir. C'est le cas aussi du test du bonhomme (voir II).

C'est aussi le cas du test de lecture d'**André Inizan** (1966) composé de 4 sous-tests, dont deux sont donnés ci-dessous en :

Test 1 :



Bébé	Bonjour
La pipe	Un cheval
La tête	Une maison
Le coq	La montagne
La table	Une cage
L'école	Un lapin
Un ami	Des dessins
Vendredi	Des marionnettes

---

<i>Bébé</i>	<i>Bonjour</i>
<i>La pipe</i>	<i>Un cheval</i>
<i>La tête</i>	<i>Une maison</i>
<i>Le coq</i>	<i>La montagne</i>
<i>La table</i>	<i>Une cage</i>
<i>L'école</i>	<i>Un lapin</i>
<i>Un ami</i>	<i>Des dessins</i>
<i>Vendredi</i>	<i>Des marionnettes</i>


Test 2 :

Je dessine papa qui fume la pipe 

---

Je colorie la carotte du lapin 

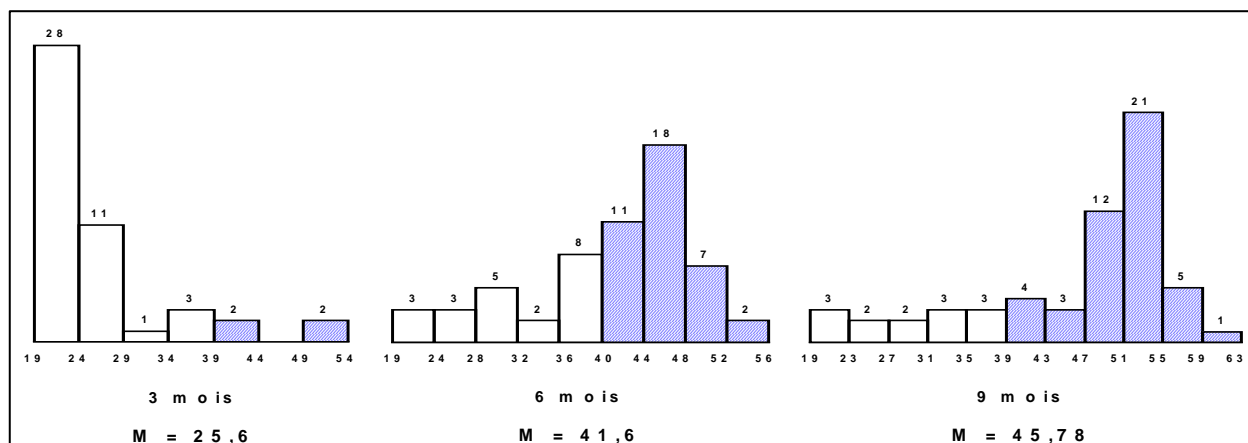
---

Une pomme est tombée de l'arbre, je la dessine 

---

Je dessine trois cerises attachées ensemble

Dans le cadre de sa recherche PREDIC, **Elise Boxus** (1969,48) a présenté les 4 sous-tests aux mêmes élèves de 1<sup>re</sup> primaire après 3, 6 et 9 mois, montrant les courbes en i, en cloche (Gauss) et en J (Leclercq & al., 1998,64).



### F3. Evaluation à notes (et performances) DEFINITIVES vs AMELIORABLES

#### a) A notes (et Performances) définitives

La plupart du temps, la notation porte sur une performance que l'étudiant ne pourra plus modifier. Ce caractère définitif culmine dans les consignes ne permettant pas les « ratures » ou dans les QCM ne permettant pas les modifications de réponses. Le SMART (Système Méthodologique d'Aide à la Réalisation de Tests) de l'université de Liège permet, par une « ligne de repentir » (re) pour la réponse à chaque question et une autre pour le degré de certitude de modifier (une fois) son avis pendant l'épreuve. Voici un extrait de formulom (Formulaire destiné à la Lecture Optique de Marques) utilisé dans le projet MOHICAN :

#### Epreuve de VOC

Q	1	2	3	4	5	6	7	Q	1	2	3	4	5	6	7	Q	1
1	☐	☐	☐	☐	☐	☐	☐	10	☐	☐	☐	☐	☐	☐	☐	19	☐
Cert	0	20	40	60	80	100		Cert	0	20	40	60	80	100		Cert	
re	☐	☐	☐	☐	☐	☐	☐	re	☐	☐	☐	☐	☐	☐	☐	re	☐
Q	1	2	3	4	5	6	7	Q	1	2	3	4	5	6	7	Q	1
2	☐	☐	☐	☐	☐	☐	☐	11	☐	☐	☐	☐	☐	☐	☐	20	☐
Cert	0	20	40	60	80	100		Cert	0	20	40	60	80	100		Cert	
re	☐	☐	☐	☐	☐	☐	☐	re	☐	☐	☐	☐	☐	☐	☐	re	☐
Q	1	2	3	4	5	6	7	Q	1	2	3	4	5	6	7	Q	1
3	☐	☐	☐	☐	☐	☐	☐	12	☐	☐	☐	☐	☐	☐	☐	21	☐
Cert	0	20	40	60	80	100		Cert	0	20	40	60	80	100		Cert	
re	☐	☐	☐	☐	☐	☐	☐	re	☐	☐	☐	☐	☐	☐	☐	re	☐

On trouvera dans Wood (1977, p. 234) une discussion de cette question. Wood, qui se base sur les travaux de Pippert (1966), Copeland (1972), Foote et Belinsky (1972), Relling et Taylor (1972), Jacobs (1974), Pascale (1974) et Lynch et Smith (1975), conclut qu'il est préférable de permettre aux étudiants de modifier leurs réponses en cours d'épreuve.

A des fins de recherche, le SMART pourrait calculer deux notes : celles SANS correction et celles AVEC correction.

Cependant, la seule note calculée est, conformément à ce qui a été promis aux étudiants, la note AVEC les corrections.

#### b) Notes (et Performances) améliorables

Cette modalité peut prendre une variété de formes :

##### Version « revue » (ou bis) du travail

Un professeur peut permettre à l'étudiant de modifier un rapport après avoir pris connaissance des critiques (écrites) faites sur la première version par le professeur. C'est ce qu'ont fait D. Leclercq et V. Jans (1999) en attribuant comme **note** finale la moyenne des deux rapports. Ce principe évite que l'étudiant remette une première version bâclée, car elle pèsera pour la moitié des points. En outre, la capacité d'un étudiant à tirer parti des critiques qui lui sont faites mérite en soi d'être prise en compte. La seule situation dramatique est celle où la première version est tellement mauvaise que sa **note** (faible) est « irrattrapable ».

##### Défense orale d'un écrit

...en vue d'améliorer sa performance (écrite) sur cette base. Le professeur peut n'octroyer ce « droit à l'amélioration » que si la copie atteint un minimum fixé au préalable. De la même façon, la défense orale fait partie intégrante de la performance « mémoire de fin d'études ». Il est fréquent que les professeurs aient fixé leur **note** pour la version écrite avant cette défense.

##### Feedback minimal

Le professeur peut se contenter de signaler une note (sans précisions) ou le nombre d'erreurs (sans dire lesquelles) avant de permettre un nouvel essai.

## G. Les méthodes d'évaluation : COMMENT ?

### G1.

#### Evaluation à source OBJECTIVE vs SUBJECTIVE

##### a) L'évaluation objective

L'objectivité n'existant pas, on la définit comme « le consensus entre experts ». En évaluation, pour y parvenir, on s'efforce souvent d'obtenir des échantillons de comportements dont la mesure puisse être la même, quel que soit le juge, à condition qu'il marque son accord sur les instruments et les critères (idéalement prédéfinis) et les applique. Des échelles d'évaluation descriptives (c.-à-d. où les échelons ou niveaux sont décrits) est d'autant plus objective que la description de chaque échelon comporte moins d'ambiguïté.

##### b) L'évaluation subjective

Dans l'allo-estimation (voir F1) le professeur, l'entraîneur du sportif, le sélectionneur de l'équipe estime la qualité de la performance d'un candidat

-soit en prédiction : avant qu'elle soit effectuée (X doit-il être sélectionné dans l'équipe ?), invérifiable si la réponse (et l'action) est NON.

-soit en post-diction : après qu'elle ait été effectuée, mais avant qu'elle soit mesurée avec des instruments précis.

Ex. : Grâce au magnétoscope, on peut revoir la performance d'un joueur de football par exemple et compter le nombre de ses « lâchers de balle », de ses « duels gagnants », et « duels perdants », de ses « tirs cadrés », etc.).

Dans l'auto-estimation l'évalué juge ses performances, ...

. Soit en prédiction (Si on m'interroge sur la géographie de la Corse, je pense réussir x % de réponses).

. Soit en post-diction (Ma réponse à la question « *Quelle est la capitale de la Corse ?* » est « *Calvi* », avec 75 % de chances d'avoir raison).

#### G2. Evaluation en contexte NATUREL vs STANDARDISE (LABO)

##### a) Evaluation en contexte naturel

Cette évaluation porte sur des « cas » naturels, non manipulés sans idée de comparaison, dans les contraintes de la vie réelle. C'est le cas d'études historiques, biographiques, d'expériences innovantes. En stage, les contextes (et circonstances) diffèrent d'un lieu de stage à l'autre.

##### b) Evaluation en contexte standardisé

Des précautions méthodologiques sont prises pour assurer la comparabilité des résultats entre groupes :

1°: ligne de base (prétest) ; expérience (intervention) ; mesure des effets (post-test) ; arrêt de l'intervention (y a-t-il retour à la base ?).

2°: Les groupes (expérimental et de contrôle) sont tirés au hasard (aléatoirement) de la même population, et suffisamment nombreux. La variable est souvent le groupe (la classe) et non l'élève isolé.

3 : Un plan de Solomon (à 4 groupes) :

	Prétest	Intervention	Post-test
Groupe A	O1	X	O2
Groupe B	O1	-	O2
Groupe C	-	X	O2
Groupe D	-	-	O2

##### c) Plans quasi expérimentaux

La plupart du temps, les observations pédagogiques portent sur les classes telles qu'elles se présentent, sur les seuls maîtres volontaires, sur les seuls élèves présents, sur les seuls parents qui ont répondu à l'invitation et sur les seuls directeurs qui ont marqué leur accord, etc. Quelle que soit la rigueur des méthodes et des instruments de mesure, la rigueur dans la comparabilité des échantillons ne pourra être assurée entre les sujets « expérimentaux » et les sujets « contrôle ». (de Landsheere, 1979, 207).

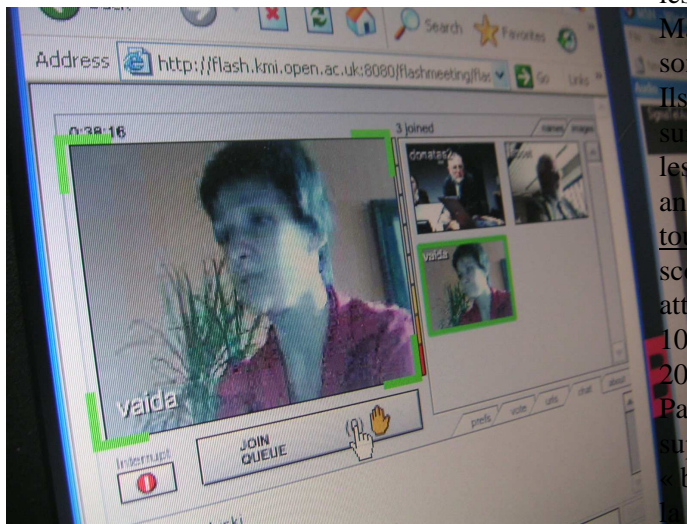
### G3. Evaluation STANDARDISEE ou ADAPTATIVE

#### a) Evaluations standardisées

Tous les étudiants d'une classe reçoivent les mêmes questions, y répondent dans les mêmes conditions (la surveillance le garantit) et sont évalués selon les mêmes critères, avec la même procédure. Les Américains parlent de « tests objectifs » quand la correction est automatisable, qu'ils ont les mêmes exigences, le même « passing score »).

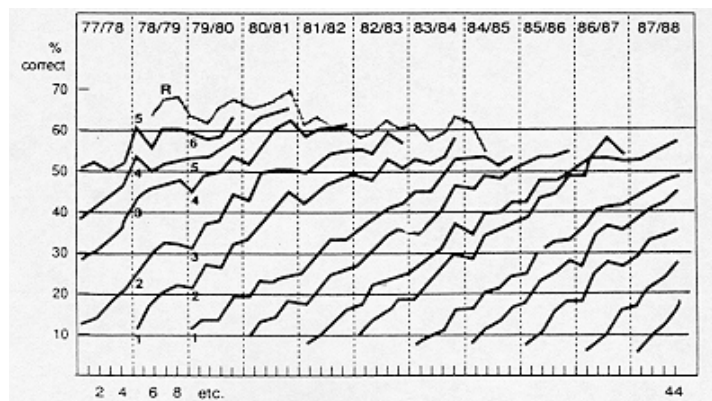
La **standardisation de la surveillance** se posant avec acuité dans les examens en EAD (Enseignement A Distance), des « *Trusted Third Parties* », Tiers de Confiance ou partenaires fiables, assurent cette surveillance. Plusieurs universités américaines ont des accords avec l'ambassade des USA dans différents pays pour que s'y déroulent des épreuves EAD sous surveillance « rapprochée ».

Voici une situation semblable, mais en 3 points) où sur l'on voit les trois sites : Liège (Belgique), Kaunas (Lithuanien) et Chicago (USA) dans le cadre d'une présentation d'un travail par une enseignante participante au DES de Pédagogie de l'Enseignement Supérieure (FORMASUP) de l'université de Liège.



#### b) Evaluations adaptatives...

...par le contenu : Permettre à chaque étudiant de choisir sa question (par exemple le thème d'une dissertation) en est un exemple. **L'examen oral** en est un autre, évoluant dans le temps, puisque l'interrogateur modifie ses questions subséquentes (leur nature et leur nombre) en fonction des réponses de l'évalué aux précédentes (principe de Wald, 1943). On peut automatiser un tel testing adaptatif par ordinateur (cf. Leclercq, 2005, Questions approfondies d'évaluation pédagogique. Editions de l'Université de Liège).



#### ...par les exigences :

Les « **Tests de progression** » pratiqués tous les 3 mois à la faculté de médecine de Maastricht (Leclercq et Vandervleuten, 1998) sont obligatoires.

Ils comportent 250 QCM Vrai-Faux, portent sur toute la médecine, et sont imposés à tous les étudiants en médecine (de la 1<sup>o</sup> à la 6<sup>o</sup> année). Le contenu est donc le même pour tous. Cependant, les seuils de réussite (passing score) diffèrent selon l'année de l'étudiant (on attend de l'étudiant de 1<sup>o</sup> année qu'il réussisse 10% de ces questions, à ceux de 2<sup>o</sup> année, 20%, etc.)

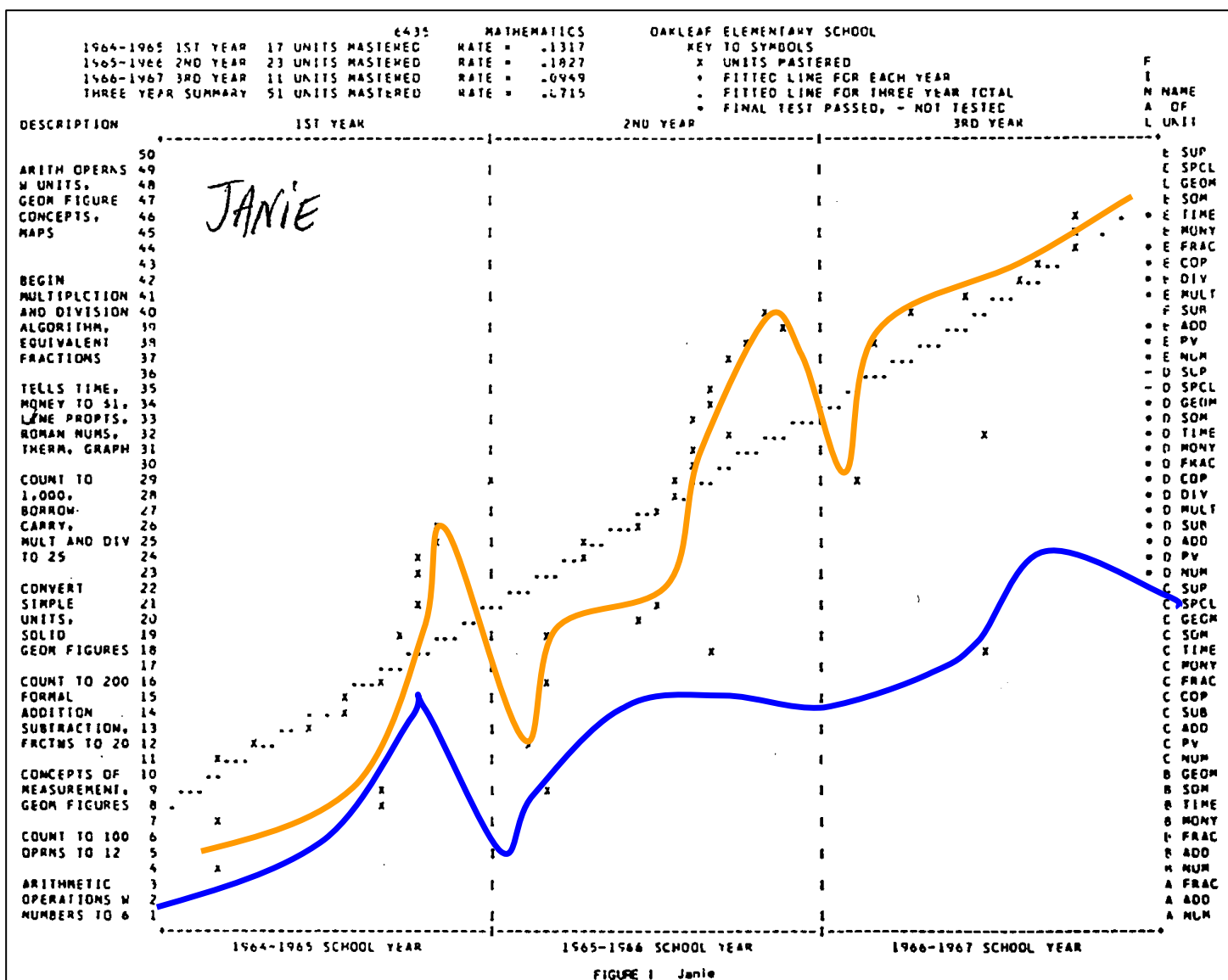
Parmi les avantages de cette méthode, la suppression du « bachotage » (on ne « bloque » pas la médecine en un week-end), la visibilité du genre de questions, et l'impossibilité d'oublier la matière des années passées, etc.

...par le moment Dans la plupart des cas envisagés jusqu'ici, l'étudiant n'a pas le choix du moment du testing. Ce n'était pas le cas dans le système IPI (Individually Prescribed Instruction) développé par le LRDC (Learning Research & Development Center) de l'université de Pittsburgh, dirigé par Robert Glaser.

Dans ce système, chaque élève avance à sa vitesse (via des livrets de travail accompagnés de cassettes sonores), puis se fait tester quand il veut (quand il estime être prêt) à l'aide d'un des terminaux d'ordinateur présent dans la classe. En fonction du résultat, le programme informatique prescrit soit de passer à une unité de contenu supérieure, soit de recommencer,

soit d'essayer une voie alternative. Ce système régule les accès inopportuns à l'ordinateur (quand l'élève n'est pas prêt) par le concept de « self reliant student » (qui sait ce gérer, un titre que l'on peut gagner, perdre, regagner, etc.

Dans ce système IPI, on observe de très grandes différences de progression entre les élèves d'un même âge comme le montre ci-dessous la progression de deux élèves en trois ans dans le cursus d'arithmétique : On trouvera plus de détails sur l'IPI dans Leclercq (2005). « Conception et Evaluation de Curriculums de Formation ». Editions de l'Université de Liège.



...Par le choix (automatisé) des questions selon les performances préalables

**a) Le modèle de RASCH (1960)<sup>1</sup>**

Ce modèle mathématique (*Item Response Theory*) dû au danois RASCH permet d'exprimer la probabilité P qu'une personne donnée de capacité C dans un domaine fournisse la réponse correcte à une question de difficulté D, sachant que C et D sont exprimés dans une même échelle, avec les mêmes unités. Un étudiant à compétence C élevée est un étudiant qui réussit des questions à difficultés D élevées.

**b) Le testing**

Au départ, l'étudiant reçoit une demi douzaine de questions dont la difficulté D est connue du testeur (qui a extrait ces questions d'une banque de questions calibrées, c'est-à-dire dont on connaît les D). Les réponses (correctes ou non) de cette première série de questions permettent de se faire une première estimation de la compétence C de l'étudiant elle aussi exprimée en Wits (par ex. la moyenne des Wits des questions réussies par lui)<sup>2</sup>

La question suivante est choisie (par l'ordinateur) selon la compétence C de l'étudiant, compétence réévaluée après chacune de ses réponses. Les questions successives d'un tel test sont choisies (automatiquement par le programme) pour que l'étudiant aie 50% de chance de réussir, étant donné ce que l'on sait de la Difficulté D de la question d'une part et de sa compétence C dans la matière d'autre part. On pourrait imaginer d'autres testings adaptatifs, moins « mathématiques » et plus « sensibles au contenu ».

Le **testing par ordinateur**, quand il impose à l'étudiant de répondre à chaque question sans pouvoir revenir en arrière **prive les étudiants** d'une possibilité stratégique (test **wiseness**) : lire et répondre aux questions dans l'ordre qu'il veut.

...Par le moment de l'arrêt de l'examen

**a) L'oral**

Fréquemment des enseignants universitaires « découvrent » que les résultats de leurs étudiants aux examens écrits se distribuent sous forme gaussienne alors que les résultats aux examens oraux se distribuent sous forme en U.

Pour la distribution gaussienne, c'est fatal si l'on choisit des questions sur base de leur difficulté (taux de réussite proche de 50%) et de leur pouvoir discriminatoire (rpbis élevé).

Pour la courbe en U de l'oral, on peut avancer l'hypothèse d'un biais de confirmation. Pour pouvoir prendre sereinement une décision (de réussite ou d'échec), le professeur a intérêt à disposer de données non ambiguës, allant, idéalement, toutes dans le même sens. Il y a donc une tendance, après que l'étudiant ait répondu à quelques questions, à en poser d'autres pour confirmer ce qu'on vient d'observer et pouvoir ainsi mettre fin plus tôt au dialogue.

**b) Le testing séquentiel de Wald ou « Quand le dialogue doit-il être arrêté ? »**

Nous avons développé ailleurs (Leclercq, 2005, Questions approfondies d'évaluation pédagogique) le modèle de Wald et son application à l'éducation.

<sup>1</sup> Voir Leclercq, 1987, chap.3

<sup>2</sup> Une réussite nulle (0%) ou totale (100%) ne permettant pas de calculer cette première estimation).



## G4. Evaluation DE PROGRESSION ou A REPERES MOBILES

### a) L'évaluation à REPERES MOBILES

La plupart du temps, le professeur évalue l'état de compétence des étudiants par rapport (en se référant) aux attentes généralement admises pour ce groupe d'âge. Ainsi, en 3<sup>e</sup> primaire, les questions sont plus difficiles qu'en 2<sup>e</sup> mais plus faciles qu'en 4<sup>e</sup> primaire, le taux de réussite restant généralement le même dans ces diverses années. Dans le test d'intelligence de BINET est considérée comme typique d'un âge donné une question réussie par 75% des enfants de cet âge.

Si un même test (de mathématique par exemple) est proposé à des élèves d'années différentes (2<sup>o</sup>, 3<sup>o</sup>, 4<sup>o</sup>, 5<sup>o</sup>, etc..) on constate

- a) que pour chaque année, il y a une dispersion des résultats (disons un écart-type de 5 points) autour de la moyenne.
- b) Qu'entre deux années successives la moyenne s'est élevée (aux USA d'un écart-type par an, c.-à-d. de 0,1 écart-type par « mois scolaire ouvrable »).

On pourrait dire que la barre monte avec les années, l'élève ayant toujours à peu près les mêmes genres de résultats.

La perfection n'est pas connue, car les années suivantes les tâches sont toujours plus difficiles ce qui compense le fait qu'on est toujours plus compétent.

C'est le cas pour le footballeur qui « monte de division ». Il avait beau être le meilleur en Division 3, il va devoir être maintenant au niveau de la Division 2.

### b) L'évaluation à repères FIXES : les tests DE PROGRESSION

Le plus bel exemple de cette méthode est le principe des *Progress Tests* développé (par Vandervleuten & Wynen, 1990) à la Faculté de médecine de Maastricht (voir Leclercq & Vandervleuten, 1998, 199 pour les détails). Les étudiants reçoivent cependant un (long) feedback sur leur position (en notes Z) par rapport aux autres étudiants de leur année.

En ski, par exemple, les « grades » sont connus à l'avance (les « flocons », les « étoiles ») et on peut en passer plusieurs la même saison si l'on veut, s'y présenter autant de fois que l'on veut jusqu'à ce que l'on réussisse, ce qui n'est pas le cas pour les *Progress Tests* à Maastricht.

Les épreuves de l'IPI à Pittsburgh constituent de tels Tests de progression (voire les contenus dans la verticale du graphique) accessibles à tout moment.

## H. L'annonce des critères de l'évaluation : AVEC QUELLE PREVISIBILITE?

### H1. Critères ANNONCES dans la consigne vs CACHES

#### a) Critères ANNONCES dans la consigne

Tous les critères sont précisés dès le départ. Par exemple, dans les QCM, la consigne précise :

-la nature de la Question (une seule solution est correcte, etc.) ;

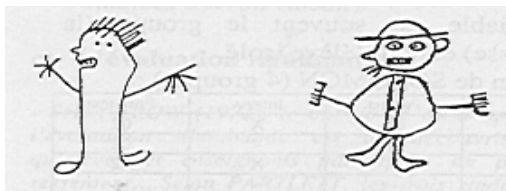
-comment répondre (par exemple en accompagnant sa réponse d'un degré de certitude).

-les conséquences en cas de Réponse Correcte (RC) ou Incorrecte (RI), en appliquant un barème de tarifs annoncé (TC = Tarif en cas de réponse correcte ; TO = en cas d'omission ; TI = en cas de Réponse Incorrecte).

Pour les performances complexes, on annonce quels critères seront appliqués.

#### b) Critères ABSENTS de la consigne ou « Faites de votre mieux ».

La consigne est « ouverte » laissant une grande liberté d'expression à celui qui répond. C'est le cas du « test du bonhomme » où l'on demande à des enfants d'âges divers (et au même enfant à des âges différents) de dessiner un bonhomme « le plus complet possible ». On applique, ensuite, une « grille » (de Goodenough) permettant de situer le degré d'évolution dans la représentation mentale du schéma corporel (présence du tronc, du cou, de la paume, du talon, etc.).



#### c) Situations intermédiaires

La dissertation en est une car on annonce souvent quelques critères (mais flous comme : « originalité des idées », etc.).

### H2. Evaluation à planification PREORDONNEE vs REPONDANTE

#### a) L'évaluation préordonnée

Parlant de l'évaluation des programmes (applicable à l'évaluation de personnes), de Landsheere (1979, p.115) déclare :

*L'évaluation ordonnée a priori (pre-ordinate) ...commence par la définition systématique des objectifs et consiste à estimer dans quelle mesure ceux-ci (et pas d'autres) sont atteints, les conclusions étant le plus souvent transmises aux responsables ou aux autres chercheurs et non aux éducateurs directement impliqués*

#### b) L'évaluation répondante

Cette expression, de Stake, désigne une démarche focalisée plus sur les processus que sur les résultats. Pour de Landsheere (1979), « une l'évaluation est dite répondante,

*-si elle porte plus sur les actions réelles que sur les intentions initiales ;*

*-si elle répond à un désir d'information de ceux dont l'action est évaluée ;*

*-si l'on conclut au succès ou à l'échec du programme éducatif en tenant compte des valeurs des individus impliqués ;*

*-si l'on fait évaluer la description que l'on fait de la situation et les résultats de l'évaluation par ceux qui sont les agents de l'action (enseignants, autorités scolaires...)*

*L'évaluation répondante se veut directement utile à des personnes particulières ; elle tient compte de leurs intérêts et utilise leur langage. (p.115)*

Des effets positifs imprévus (*side effects*, obliques ou seconds) sont parfois assez importants pour justifier la poursuite d'un programme qui n'atteint cependant pas les résultats escomptés. A l'opposé, les effets négatifs peuvent revêtir une telle gravité que l'arrêt du programme, par ailleurs réussi, s'impose. C'est pourquoi Scriven propose qu'en certains cas les évaluateurs analysent les situations en toute indépendance et choisissent de mesurer les effets qu'ils croient observer (*goal free evaluation*) plutôt que de se laisser guider par les objectifs explicites du programme (*goal based evaluation*). » (p.114)

# I. La transparence de l'évaluation : AVEC QUELLE VISIBILITE ?

## I1. Le modèle

Dans « L'évaluation en formation », Barbier (1990) distingue trois étapes : le processus d'évaluation, le jugement et la décision. Il insiste sur leur caractère explicite ou implicite. On remarquera la proximité du modèle de Barbier avec celui d'un procès en justice : le déroulement du procès (processus), le jugement par le jury (ex : Coupable, avec circonstances atténuantes), la décision par le juge (ex : 5 ans de prison avec sursis). Certains aspects sont publics (obligatoirement chez nous), d'autres à huis clos.

Demarteau (1998) a systématisé cette approche. Dans ses schémas (ci-après), il situe l'implicite en bas et en grisé et l'explicite en haut et en clair. Cette grille, appelée APEP (Analyse des Pratiques en Evaluation de Programmes) se présente comme suit dans le cas d'une évaluation totalement implicite :

T1 = L'évaluation totalement implicite

Processus	Jugement de valeur	Décision	Explicite
			Implicite

### Exemple :

Un employé a été « surveillé » à son insu (**processus**), et sur base d'un rapport (oral et confidentiel) d'un informateur (l'évalué ne saura jamais qui, quand ni comment), son patron s'est fait une idée négative (**jugement de valeur**) de cet agent, et s'est juré (**décision**) en son for intérieur de ne plus lui confier de responsabilité.

L'agent ne saura jamais qu'il a été observé, continuera à croire que son patron l'apprécie et attribuera à des raisons techniques sa non affectation à un poste de responsabilité.

## I2. A processus

### EXPLICITE ou IMPLICITE

#### a) Processus explicite

On dispose d'un « prescrit procédural ». C'est le cas pour les examens organisés par des instances officielles (comme le SELOR en Belgique).

#### Exemple

L'évaluation des universités, organisée par l'Agence Qualité de l'enseignement supérieur de la CFWB ou, sur demande, par l'European Universities Association (EUA), se déroule comme suit :

1. L'université évaluée procède à un rapport d'auto-évaluation
2. Les experts externes prennent connaissance de ce rapport
3. Les experts font une visite (plusieurs jours) sur le terrain où ils examinent les documents et interviewent des personnes de leur choix
4. Les experts transmettent à l'université un pré-rapport
5. Une rencontre avec les autorités de l'université est organisée avec les experts qui présentent leur pré-rapport
6. L'université évaluée réagit au pré-rapport
7. Les experts remettent leur rapport définitif.

A ce moment, on se trouve dans la situation suivante :

Processus	Jugement de valeur	Décision	Explicite
			Implicite

Car l'université n'a pas encore pris de décision et qu'il ne rentre pas dans les attributions des experts externes d'en prendre.

#### b) le processus implicite

Voir l'exemple de l'employé en J1.

### 13. À jugement EXPLICITE ou IMPLICITE

#### a) Jugement explicite

L'exemple précédent en constituait déjà un exemple. Toutes les procédures débouchant sur un score connu en constituent d'autres.

#### b) Jugement implicite

**Exemple**

A la suite d'un entretien d'embauche, le candidat connaît la procédure (qu'il vient de vivre), connaît la décision (non retenu) mais, souvent, ne connaît pas le jugement (pourquoi a-t-il été écarté ? quels sont les qualités de ses rivaux qui lui sont supérieures ?).

On est dans la situation :

Processus	Décision	Explicite
Jugement de valeur		Implicite

**Exemple**

La législation belge prévoit, dans le règlement des examens universitaires, que l'étudiant a le droit de consulter sa copie notée par le professeur dans un certain délai (par exemple 15 jours après communication de la note). Il s'agit là d'une procédure visant à rendre explicite le jugement (et ses détails) là où c'est ressenti comme nécessaire par l'évalué. Le droit à consulter la copie d'autrui est moins réglementé.

On a ainsi permis de rendre explicite le jugement :

Processus	↑	Décision	Explicite
Jugement de valeur			Implicite

### 14. Le caractère EXPLICITE ou IMPLICITE de la DECISION

La notion de sursis est un exemple : il n'y aura pas d'effet (par exemple pas d'emprisonnement), mais à condition qu'il n'y ait pas récidive.

#### a) Décision explicite

Quand quelqu'un le dit tout haut (le sursis), on est dans la situation suivante :

Jugement de valeur	Décision	Explicite
		Implicite

Car une non action est une décision.

#### b) Décision implicite

On peut décider (mais sans le dire) que « si cela se reproduit encore, alors... ». On est dans la situation suivante :

Jugement de valeur	Décision	Explicite
		Implicite

Bibliographie en fin de volume.