APPORTS DES MESURES METACOGNITIVES LORS D'UN TEST DE SELECTION PORTANT SUR LA COMPREHENSION D'UN ARTICLE SCIENTIFIQUE EN 1ERE CANDIDATURE DE LA FACULTE DE MEDECINE

Jean-Luc Gilles[1]

Centre d'Auto-Formation et d'Evaluation Interactives Multimédias de la
Faculté de Psychologie et des Sciences de l'Education de l'Université de Liège

INTRODUCTION

Depuis l'année académique 1994-1995, le Centre d'Auto-Formation et d'Evaluation Interactives Multimédias (CAFEIM) de la Faculté de Psychologie et des Sciences de l'Education (FAPSE) de l'Université de Liège (ULg) propose aux enseignants un Service Méthodologique d'Aide à la Réalisation de Tests (SMART). Les procédures d'évaluation proposées permettent un questionnement ayant recours à trois types de dispositifs technologiques :

- la Lecture Optique de Marques (LOM) où les étudiants cochent leurs réponses en amphithéâtre sur des *formuloms*[2] ;
- le testing informatisé interactif où les réponses sont fournies à l'aide des ordinateurs du CAFEM (sous surveillance dans le cas des évaluations certificatives) ou situés sur d'autres sites et reliés à notre serveur grâce au réseau Intranet de l'Ulg ou à l'Internet (GILLES, 1998) ;
- les boîtiers de vote électronique (depuis cette année académique 1998-1999) qui permettent de tester les acquis d'un grand groupe d'étudiants en amphithéâtre et de fournir en temps réel un feedback à l'auditoire.

Des mesures gouvernementales visant à fixer le nombre de médecins pratiquants imposent à la Faculté de Médecine une sélection des étudiants en fin de 3ème candidature dès l'an 2000. Un des tests de sélection, réalisé en collaboration[3] avec le CAFEIM, porte sur la Compréhension d'un Texte Scientifique (test CTS). Ce test a été soumis pour la première fois aux étudiants de première candidature en médecine en mai 1998.

1. Vers une gestion de la qualité des évaluations certificatives dans le cadre du SMART

La réalisation d'une évaluation certificative, peut être schématisée dans un cycle à huit étapes (GILLES & LECLERCQ, 1995) (voir figure 1). Une série de recommandations en vue d'augmenter la validité et la fidélité des examens sont présentées dans les cadres rectangulaires ombrés qui entourent le schéma. Les recommandations soulignées en continu sont systématiquement proposées dans les évaluations ayant recours au SMART. Celles qui sont soulignées en pointillés sont, à notre avis, encore trop peu suivies par les enseignants, certaines ne le sont pas du tout et ne sont pas soulignées dans les cadres. Les plages hachurées délimitent les champs d'action de logiciels d'aide à la réalisation de l'étape concernée.

Nous travaillons actuellement à l'intégration de ce cycle dans un système qualité en nous basant sur les recommandations de la norme internationale ISO 9004-2[4] « *Gestion de la qualité et éléments de système qualité - Lignes directrices pour les services.*

[1] L'auteur tient à remercier le Professeur Dieudonné Leclercq pour son soutien et ses précieux conseils.

[2] *Formuloms* signifie « formulaires destinés à la lecture optique de marques ».

[3] Equipe de réalisation du test de compréhension d'un texte scientifique :
- Faculté de Médecine : C. Balthazart (Coordinatrice) et F. Pasleau ;
- Centre d'Auto-Formation et d'Evaluation Interactives Multimédias de la FAPSE : Prof. D. Leclercq et J.-L. Gilles.

[4] *Gestion de la qualité et éléments de système qualité – Partie 2 : Lignes directrices pour les services*, Organisation internationale de normalisation, Case postale 56, CH-1211 Genève 20, Suisse. Numéro de référence : ISO 9004-2 :1991(F). Première édition 1991-08-01, corrigée et réimprimée 1993-05-01.

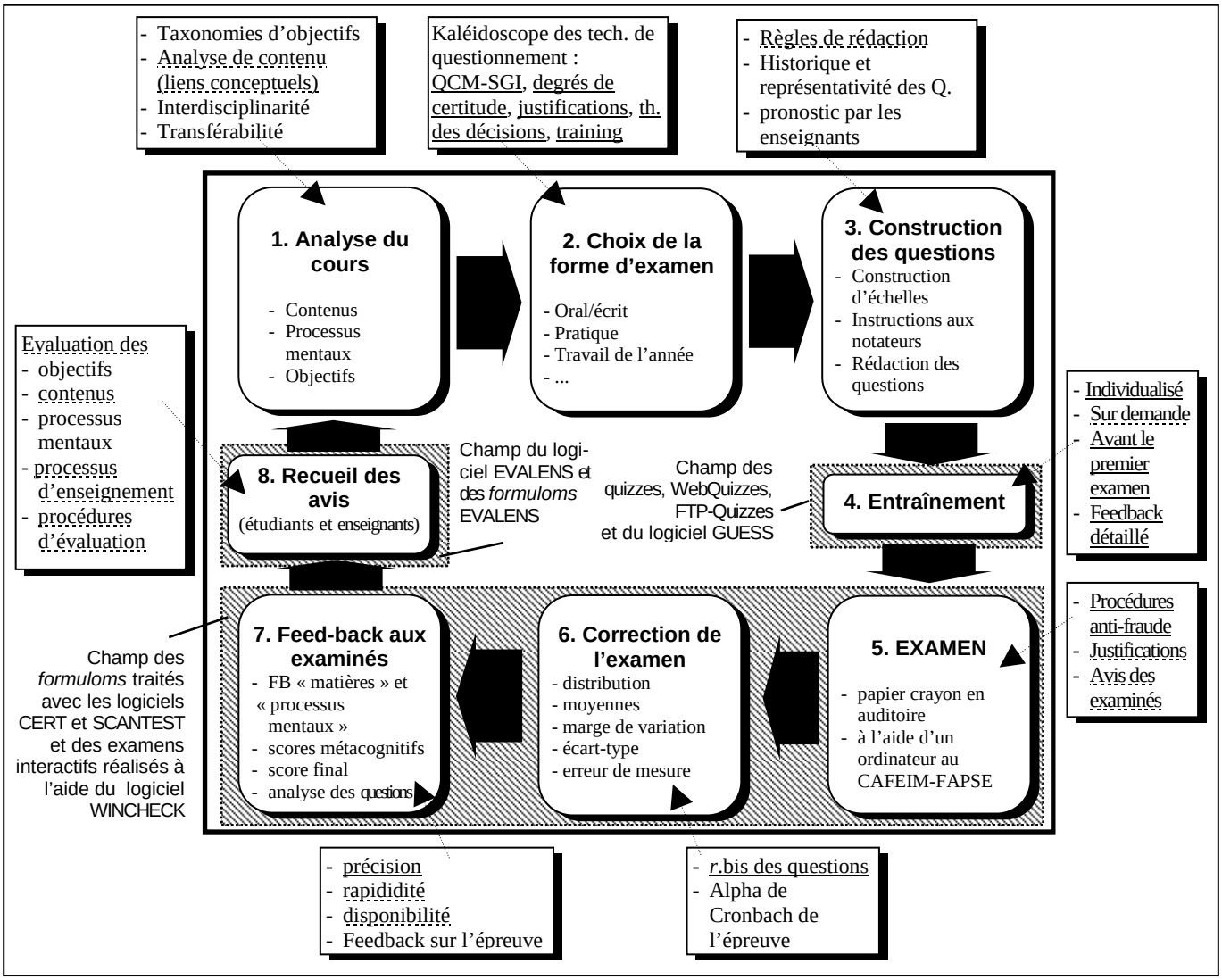

Figure 1 : cycle de réalisation d'une évaluation certificative dans le cadre du SMART

L'utilisation des degrés de certitude fait faire un bond qualitatif (LECLERCQ et GILLES, 1995) aux QCM ou aux QROC à condition qu'une série de règles méthodologiques soient respectées : (1) Une consigne « probabiliste », (2) un barème de tarifs calculé selon la théorie des décisions, (3) le calcul d'indices de réalisme, (4) un entraînement à la procédure.

De nombreuses façons d'exprimer le degré de certitude ont été décrites (voir LECLERCQ, 1993, pp. 114-131). Selon la règle 1, seules celles où la consigne est probabiliste (non pas " peu sûr " mais " certitude comprise entre 25 et 50 % ") et où le barème des tarifs avantage une expression sans biais de la certitude (règle 2) sont considérées comme " *Admissible Probability Measurement Procedures* " par SHUFFORD & al., (1966). Le tableau 1 présente le barème des tarifs préconisé par LECLERCQ (1983, 1993)

Si vous considérez que votre réponse a une probabilité d'être correcte comprise entre	Ecrivez	Vous obtiendrez les points suivants en cas de réponse	
		Correcte	Incorrecte
0 % et 25 %	0	+ 13	+ 4
25 % et 50 %	1	+ 16	+ 3
50 % et 70 %	2	+ 17	+ 2
70 % et 85 %	3	+ 18	+ 0
85 % et 95 %	4	+ 19	- 6
95 % et 100 %	5	+ 20	- 20

Tableau 1 : barème des tarifs liés aux degrés de certitude de D. LECLERCQ

Dans le cycle de réalisation des évaluations certificatives (voir figure 1), une des activités clé de l'étape « 6. Correction de l'examen », consiste à mesurer la qualité des questions. C'est dans cette perspective que nous avons tenté de développer de nouvelles méthodes permettant d'évaluer la qualité des questions en ayant recours à l'information livrée par les degrés de certitude dans le cadre des QCM et QROC. Pour faciliter les calculs et accélérer les traitements, nous avons réalisé un logiciel intitulé SCANTEST[5] (voir figure 2). Ce programme utilise les fichiers des réponses des étudiants et de paramétrage des

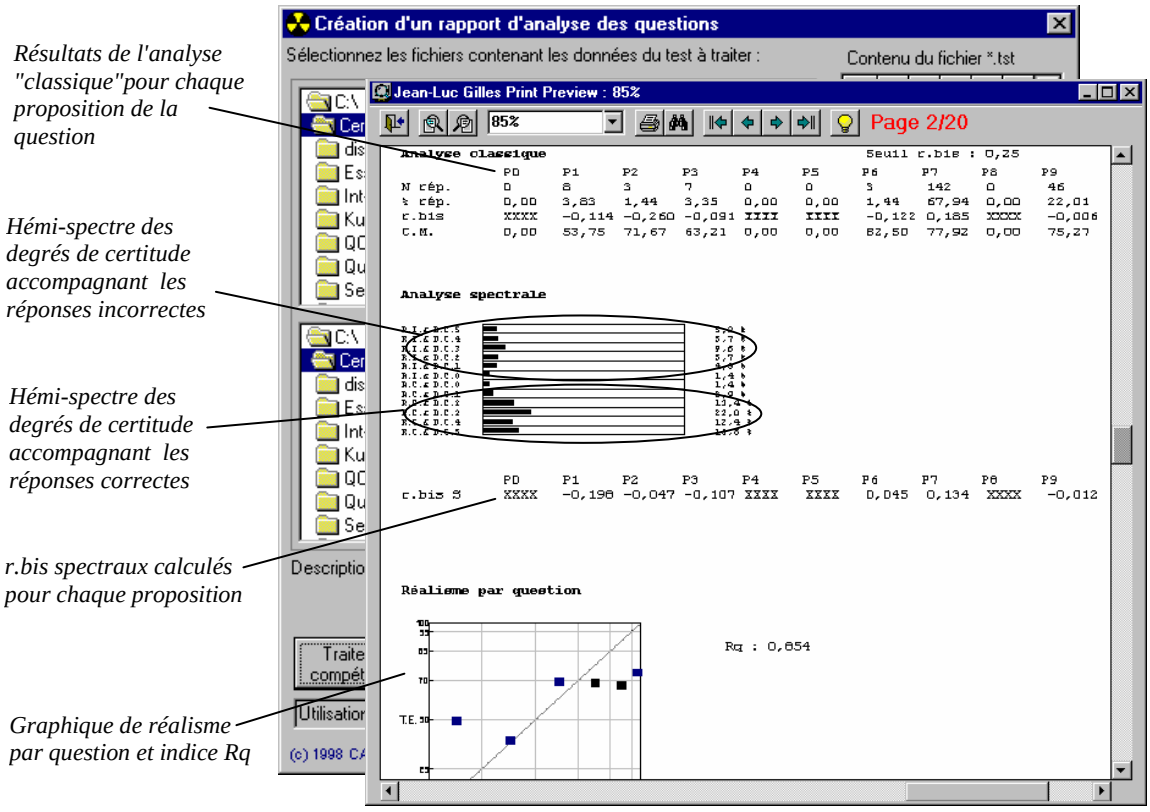

Figure 2 : exemple de rapport de traitement pour une question - logiciel SCANTEST

tests habituellement employés dans le cadre des traitements effectués à l'aide du logiciel CERT, ce qui permet de l'intégrer dans la chaîne de Lecture Optique de Marque (LOM) du CAFEIM. Les sorties peuvent être visualisées à l'écran ou imprimées. Le rapport final comprend pour chaque question :

- une analyse "classique" (ces résultats étaient déjà disponibles dans CERT) des propositions : nombre d'étudiants (N rép.) qui ont choisi la proposition, pourcentage (% rép.), corrélation bisériale de point (*r*.bis[6]) et certitude moyenne (C.M.);
- une analyse spectrale avec le graphique du profil spectral qui reprend pour chaque degré de certitude le pourcentage des étudiants qui ont choisi une proposition incorrecte (hémi-spectre des réponses incorrectes accompagnées des degrés de certitude 5, 4, 3, 2, 1, 0) et le pourcentage des étudiants qui ont répondu correctement (hémi-spectre des réponses correctes accompagnées des degrés de certitude) ;
- les *r*.bis spectraux (*r*.bis S) calculés en corrélant les degrés de certitude qui ont accompagné les réponses avec les choix/rejets des différentes propositions de la QCM;
- un graphique de Réalisme par question (Rq) reprenant les taux d'exactitude pour chaque degré de certitude accompagné du Rq calculé à l'aide de la formule décrite par GILLES (1997).

Des synthèses sont également proposées :

- un récapitulatif des certitudes moyennes pour les réponses correctes et incorrectes accompagné d'un graphique de cohérence métacognitive (voir figure 7) ;
- un récapitulatif des *r*.bis "classiques" pour les réponses correctes et incorrectes avec un graphique de qualité des *r*.bis reprenant ces données pour l'ensemble des questions;
- un récapitulatif des *r*.bis spectraux (r.bis S) pour les réponses correctes et incorrectes accompagné d'un graphique de qualité des *r*.bis S (voir figure 8).

[5] Le logiciel SCANTEST a été conçu et programmé par Jean-Luc GILLES dans le langage *Microsoft Visual Basic 5.0.*

[6] Le coéfficient de corrélation bisériale de point (*r*.bis) est utilisé comme indice de corrélation entre le choix des propositions d'une QCM et les scores globaux de l'épreuve. Lorsqu'une QCM "fonctionne" bien, on s'attend à un *r*.bis positif pour la réponse correcte et négatif pour la réponse incorrecte. Le seuil à partir duquel le *r*.bis d'une réponse correcte peut être considéré comme satisfaisant, est calculé en fonction du nombre de questions de l'épreuve (1/raçine du nombre de questions), moins il y a de questions plus le seuil est élevé.

2. Description du test de Compréhension d'un Texte Scientifique (CTS)

Depuis l'année académique 1996-1997, le CAFEIM collabore avec la Faculté de Médecine de l'ULG dans le cadre de la réalisation du test CTS destiné aux étudiants de 1ère candidature de cette faculté. Cette épreuve eut lieu en mars 1998 et fut précédée par un entraînement qui débuta dès octobre 1997. Le texte du test CTS de mars 1998 était tiré d'une revue de vulgarisation scientifique. Les questions étaient à choix multiple avec Solutions Générales Implicites (SGI)[7]. Sur 26 QCM, 17 étaient de type SGI (la réponse attendue est une SGI) et 6 étaient habituelles (la réponse correcte était dactylographiée). Chaque réponse devait être accompagnée d'un degré de certitude. Toutes les réponses figuraient dans le texte. Une première simulation en grandeur réelle eut lieu en mars 1997 avec un groupe d'étudiants volontaires (N=130) de 1ère candidature de l'année académique 1996-1998. Le test utilisé à cette occasion fut ensuite proposé dans le cadre de la procédure d'entraînement (*cfr* étape 4 du cycle SMART - figure 1) aux étudiants de l'année académique 1997-1998, ceux pour qui le test final serait sanctionnant. MELON et GILLES (1998) ont procédé à une comparaison des résultats au test CTS de mars 1998 avec un test de Maîtrise du Français (MF) comportant 80 questions réparties en 4 sous-tests : orthographe (ORTH), vocabulaire (VOCA), syntaxe (SYNT) et compréhension (COMP) de 20 questions chacun. Les résultats à ces tests ont aussi été comparés avec les performances académiques des étudiants aux examens partiels de janvier. Les corrélations obtenues sont présentées dans le tableau 2. Remarquons que ce sont les performances au test CTS qui sont les plus corrélées (r = .55) avec les résultats obtenus aux partiels de janvier. Au sein du test CTS, les performances aux 17 questions dont la réponse attendue était une SGI sont encore plus corrélées avec les partiels (r = .58).

Corrélations significatives marquées en gras à p < ,05 (N=199)

		CTS		MF					Partiels
		17 Q. SGI	9 Q. HABI	80 Q.	20 Q. ORTH	20 Q. VOCA	20 Q. SYNT	20 Q. COMP	1 à 5 tests
CTS	26 Q.	**,91**	**,75**	**,60**	**,35**	**,52**	**,59**	**,48**	**,55**
	17 Q. SGI		**,41**	**,54**	**,29**	**,50**	**,50**	**,44**	**,58**
	9 Q. HABI			**,47**	**,31**	**,34**	**,50**	**,36**	**,29**
MF	80 Q.				**,75**	**,83**	**,88**	**,73**	**,39**
	20Q.ORTH					**,42**	**,54**	**,39**	**,23**
	20Q. VOCA						**,69**	**,48**	**,38**
	20Q. SYNT							**,57**	**,33**
	20Q. COMP								**,30**

Tableau 2 : corrélations des scores aux partiels et des scores avec degrés de certitude aux tests MF & CTS

3. Analyse spectrale des questions

3.1 Comparaison des profils des questions du test d'entraînement : mars '97 *vs* octobre '97

Rappelons que le test d'entraînement fut soumis une première fois en mars 1997 à un premier groupe d'étudiants de 1ère candidature (N=199) et une seconde fois en octobre 1997 à un second groupe de 1ère candidature (de l'année académique suivante, N=125). La figure 3 montre les profil spectraux de la question 1 remarquablement similaires pour les deux groupes. Faute de place, nous n'avons pu reproduire ici

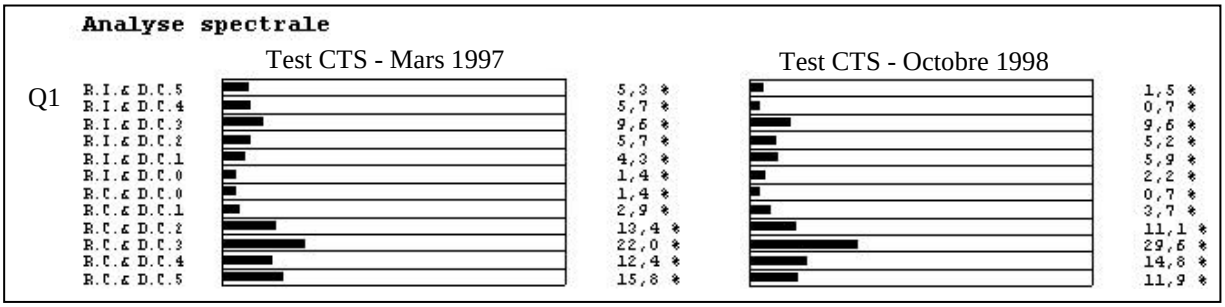

Figure 3 : comparaison des profils spectraux de la question 1 du test CTS d'entraînement (1996-1997 et 1997-1998)

les 30 autres profils des questions. Tous sont très proches, sauf celui de la question 12.

[7] Les SGI (D. LECLERCQ, 1986) autorisent, en plus des solutions habituellement proposées, les quatre possibilités suivantes : Rejet (aucune solution proposée est correcte), Toutes (toutes sont correctes), Manque (il manque des données dans l'énoncé *pour que l'on puisse choisir UNE solution comme correcte*), Absurdité (il y a une contrevérité dans l'énoncé *à dénoncer en priorité !*).

Le tableau 3 reprend les corrélations des pourcentages liés aux profils spectraux des 16 questions du test d'entraînement lorsqu'on compare les résultats obtenus lors de l'année académique 1996-1997 avec ceux de 1997-1998 : on remarque des corrélations très élevées pour la plupart des questions, sauf pour la 12 (r = 0,52). La moyenne des corrélations des 16 questions vaut 0,83.

Questions	r
1	0,92
2	0,73
3	0,98
4	0,75
5	0,79
6	0,80
7	0,94
8	0,91
9	0,91
10	0,99
11	0,65
12	0,52
13	0,90
14	0,87
15	0,95
16	0,70
Moy.	0,83
ET.	0,13

Tableau 3 : r CTS entraînement 1997 vs.1998

3.2 Profils spectraux types

Le profil spectral de la répartition des degrés de certitude peut être corrélé avec des profils type.

- Facilité Maximale Ressentie (FMR). Ce profil spectral correspond à la situation où tous les étudiants répondent correctement (Facilité Maximale...) avec un degré de certitude systématiquement maximal (...Ressentie).

- Difficulté Maximale Ressentie (DMR). Il s'agit de la situation opposée où tous les étudiants répondent incorrectement (Difficulté Maximale...) avec un degré de certitude systématiquement minimal (...Ressentie).

- Facilité Maximale Non Ressentie (FMNR). Correspond à la situation où tous les étudiants répondent correctement (Facilité Maximale...) avec un degré de certitude systématiquement minimal (...Non Ressentie).

- Difficulté Maximale Non Ressentie (DMNR). Situation où tous les étudiants répondent incorrectement (Difficulté Maximale...) avec un degré de certitude systématiquement maximal (...Non Ressentie).

- Hémi-spectre Linéaire des Réponses Incorrectes (HLRI). Ce profil correspond à la situation où les réponses incorrectes sont réparties selon une progression arithmétique dont la raison est égale à 4,8% du pourcentage de réponses incorrectes. Le pourcentage de réponses incorrectes pour un degré de certitude donné procède ainsi du précédent par addition d'un nombre constant (raison). Ce nombre constant dépend du pourcentage de Réponses Incorrectes (%RI) pour la question. Par exemple, si %RI est de 32%, la constante correspond à 0,32 x 0,048 = 0,01536 ou 1,54% , c'est le pourcentage de réponses incorrectes pour le degré de certitude 5. A partir de cette constante liée à la question, on calcule les autres pourcentages : %RI pour le degré de certitude 4 = 0,01536 + 0,01536 = 0,03072 = 3,07%, etc.

- Hémi-spectre Linéaire des Réponses Correctes (HLRC). Cette fois, ce sont les réponses correctes qui sont réparties selon une progression arithmétique dont la raison est égale à 4,8% du pourcentage de réponses correctes. Le principe de calcul des pourcentages d'utilisation des degrés de certitude pour les réponses correctes est le même que le précédent. Par exemple, si %RC est de 68% , la constante correspond à 0,68 x 0,048 = 0,03264 ou 3,26% , c'est le pourcentage de réponses correctes pour le degré de certitude 0. A partir de cette constante liée à la question, on calcule les autres pourcentages : pour le degré de certitude 1 = 0,03264 + 0,03264 = 0,06528 = 6,52%, etc.

- Hémi-spectre Géométrique des Réponses Incorrectes (HGRI). Ce profil correspond à la situation où les utilisations des degrés de certitude liés aux réponses incorrectes sont réparties selon une progression géométrique de raison 2. Le pourcentage d'utilisation du degré de certitude 5 vaut 1,6% du pourcentage de réponses incorrectes. Pour le degré de certitude 4, il vaut 1,6% + 1,6% = 32% du pourcentage de réponses incorrectes, pour le degré de certitude 3, il vaut 32% + 32% = 64% du pourcentage de réponses incorrectes, etc.

- Hémi-spectre Géométrique des Réponses Correctes (HGRC). Les utilisation des degrés de certitude liés aux réponses correctes sont réparties selon une progression géométrique de raison 2. Même principe de calcul : pourcentage d'utilisation du degré de certitude 0 vaut 1,6% du pourcentage de réponses correctes. Degré de certitude 1 : 1,6% + 1,6% = 32% du pourcentage de réponses correctes. Degré de certitude 3 : 32% + 32% = 64% du pourcentage de réponses correctes, etc.

3.3 Corrélation entre le profil d'une question et des profils type

La corrélation du profil spectral de chaque question du test d'entraînement envisagé plus haut avec les profils spectraux théoriques peut nous donner une idée de la distance qui les sépare.

Question 1		Profils théoriques							
		Facilité Maximale Ressentie **FMR**	Difficulté Maximale Ressentie **DMR**	Facilité Maximale Non Ressentie **FMNR**	Difficulté Maximale Non Ressentie **DMNR**	Hémi-spectre Linéaire Réponses Incorrectes **HLRI**	Hémi-spectre Linéaire Réponses Correctes **HLRC**	Hémi-spectre Géométrique Réponses Incorrectes **HGRI**	Hémi-spectre Géométrique Réponses Correctes **HGRC**
RI & DC5	5,3 %	0 %	0 %	0 %	100 %	1,5 %		0,5 %	
RI & DC4	5,7 %	0 %	0 %	0 %	0 %	3,0 %		1,0 %	
RI & DC3	9,6 %	0 %	0 %	0 %	0 %	4,6 %		2,0 %	
RI & DC2	5,7 %	0 %	0 %	0 %	0 %	6,1 %		4,1 %	
RI & DC1	4,3 %	0 %	0 %	0 %	0 %	7,6 %		8,1 %	
RI & DC0	1,4 %	0 %	100 %	0 %	0 %	9,2 %		16, 3 %	
RC & DC0	1,4 %	0 %	0 %	100 %	0 %		3,3 %		1,1 %
RC & DC1	2,9 %	0 %	0 %	0 %	0 %		6,5 %		2,2 %
RC & DC2	13,4 %	0 %	0 %	0 %	0 %		9,7 %		4,3 %
RC & DC3	22,0 %	0 %	0 %	0 %	0 %		12,9 %		8,6 %
RC & DC4	12,4 %	0 %	0 %	0 %	0 %		16,2 %		17,2 %
RC & DC5	15,8 %	100 %	0 %	0 %	0 %		19,4 %		34,5 %
	r	**0,37**	**-0,34**	**-0,34**	**-0,15**	**-0,56**	**0,74**	**-0,77**	**0,49**

Tableau 4 : corrélations du profil spectral de la question 1 du test CTS d'entraînement (03/97) avec les profils spectraux théoriques

Dans le cas de la question 1 posée en mars 1997, on constate des corrélations positives (voir la dernière ligne du tableau 4) avec les profils théoriques relatifs aux hémi-spectres des réponses correctes : *linéaire* (r = 0,74) et *géométrique* (r = 0,49) ainsi qu'avec le profil théorique de *facilité maximale ressentie* (r = 0,37)

.

3.4 Les ingénogrammes des profils spectraux par question

La présentation de ces corrélations sur des ingénogrammes, graphiques polygonaux à coordonnées polaires[8], permet une visualisation du rapprochement/éloignement aux profils théoriques et peut nous aider à catégoriser les questions. Cette présentation graphique peut également être utile pour comparer les performances liées à une question précise, par exemple la question 12 du test d'entraînement (voir figure 4), celle qui obtenait la moins bonne corrélation (0,52). En comparant les deux ingénogrammes de la question 12, on constate que les différences se situent essentiellement au niveau FMR (moins corrélé en 1998) et du DMNR (nettement plus corrélé en 1998). Ce type d'approche typologique pourrait se révéler particulièrement utile lorsqu'il s'agit de sélectionner des questions en vue de créer des tests équivalents.

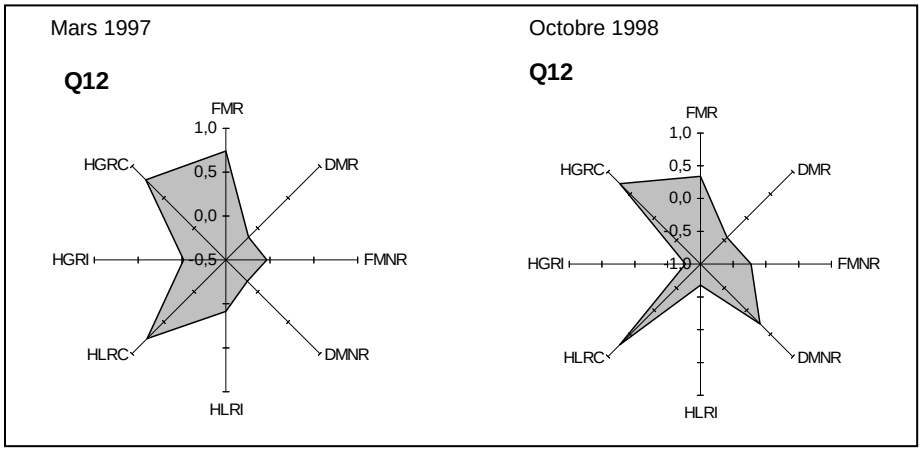

Figure 4 : comparaison profils Q12-03/97 et Q12-10/98

[8] *Les surfaces grisées ont pour but de faciliter la visualisation des profils, elles n'ont pas de valeur informative (elles sont dépendantes des angles formés avec les autres dimensions envisagées), ce sont les points d'intersection des lignes qui relient les axes qui représentent les valeurs des corrélations.*

4. Principes de *turbo-analyse spectrale*

L'analyse des Certitudes Moyennes des Réponses Correctes par question (CMRCq) et Incorrectes par question (CMRIq) ainsi que des corrélations bisériales de point spectrales (*r*.bis S) est dépendante du réalisme des sujets qui ont été testés. C'est pourquoi nous proposons de procéder à une *turbo-analyse spectrale* qui consiste à sélectionner les étudiants les plus réalistes (qui obtiennent les meilleurs scores à l'indice de réalisme[9]) pour ensuite analyser leurs seules réponses dans le logiciel SCANTEST.

4.1 Méthode

Dans le cadre de la comparaison des analyses spectrales du test d'entraînement, les groupes d'étudiants réalistes ont été constitués en établissant un classement à l'aide de l'indice de réalisme (Re[10]) puis en sélectionnant les étudiants qui obtiennent un score supérieur à 0,81. Le seuil fixé à 0,81 correspond au score à partir duquel le réalisme d'un sujet est considérée comme bon si on le compare aux performances des étudiants du 1er cycle de la FAPSE, population à partir de laquelle nous avons établi les normes (GILLES, 1996).

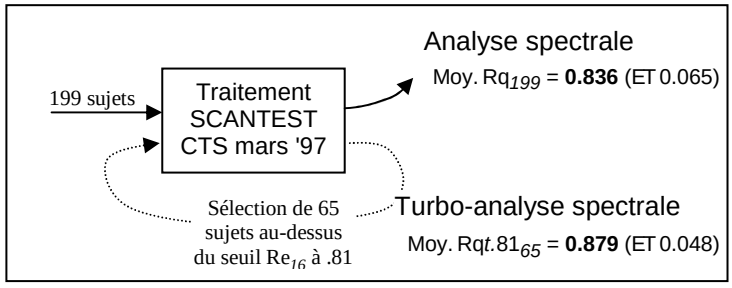

Figure 5 : exemple de turbo-analyse spectrale réalisée à partir du test CTS d'entraînement de mars 1997

4.2 Ses effets

L'amélioration du réalisme des étudiants se traduit, en principe, par de meilleurs scores à l'indice de Réalisme par question (Rq). Cet indice peut être considéré comme un indicateur de performance de la *turbo-analyse spectrale*. Nous proposons de noter l'indice "Rq*t*" (*t* étant la référence à "*turbo*"), d'y ajouter la référence au seuil Re utilisé pour la sélection des étudiants réalistes (0,81 dans notre exemple) et de compléter la notation par un indice correspondant au nombre d'étudiants dans le groupe sélectionné. Dans l'exemple relatif au test CTS d'entraînement qui s'est déroulé en mars 1997, ont constate une amélioration de 0,043 (0,879 – 0,836) de la moyenne des Rq après la *turbo-analyse spectrale*. L'amélioration se visualise pour la plupart des questions (dans 12 cas sur 16) sur les graphiques de réalisme par question. L'exemple de la question 2 permet de visualiser le meilleur alignement des Taux d'Exactitude (T.E.) aux différentes valeurs centrales des Probabilités Subjectives de Réussites (P.S.R.) liées aux degrés de certitude après la *turbo-analyse spectrale*.

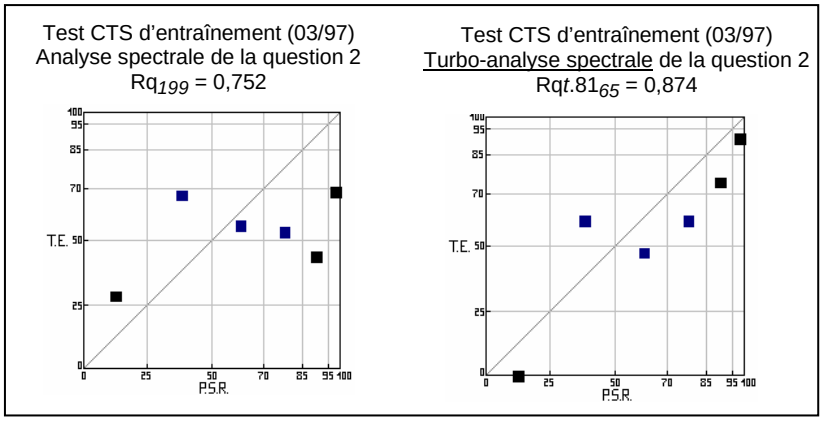

Figure 6 : comparaison du réalisme par question avant et après turbo-analyse spectrale

[9] L'indice de réalisme est calculé dans le cadre de cette étude à l'aide de la formule adaptée par GILLES (1997) de façon à ce que la performance minimale à l'indice de réalisme (par exemple pour un sujet qui se trompe systématiquement en donnant le degré de certitude maximum) soit égale à 0 et que la performance maximale soit égale à 1.

[10] Nous proposons (1) de différencier le réalisme *par étudiant* (calculé pour un étudiant à partir des réponses de ce dernier aux questions) du réalisme *par question* (calculé pour une question à partir des réponses des étudiants à la question) en attribuant le symbole Re au réalisme par étudiant et Rq au réalisme par question et (2) d'ajouter en indice, dans le cas du Re, le nombre de questions qui ont permis de calculer le réalisme, et, dans le cas du Rq, le nombre de sujets.

5. Comparaison des Certitudes Moyennes par question pour la Réponse Correcte (CMRCq) et pour les Réponses Incorrectes (CMRIq) lors du test sanctionnant

La cohérence métacognitive (réponses correctes accompagnées de certitudes élevées et réponses incorrectes de certitudes faibles) liée aux questions et calculée à partir des sujets les plus réalistes (*turbo-analyse spectrale* sur la base d'un seuil Re fixé à 0,81), apparaît dans le graphique de la figure 7. Pour chaque

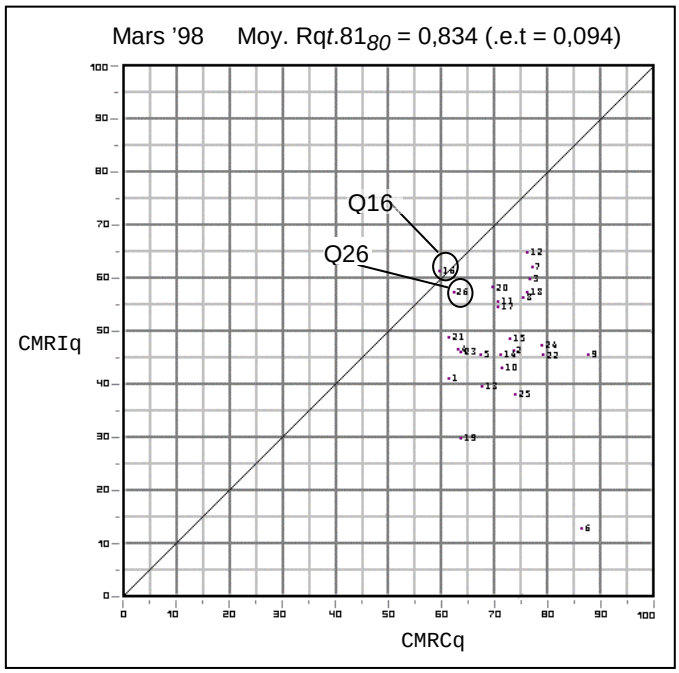

Figure 7 : graphiques des certitudes moyennes des 26 questions (réponses correctes et incorrectes) du test CTS de mars 1998

Q	CMRCq	CMRIq	Ecarts CM
6	0,867	0,125	0,742
9	0,880	0,454	0,426
25	0,743	0,378	0,365
19	0,640	0,295	0,345
22	0,796	0,452	0,344
24	0,793	0,471	0,322
10	0,717	0,429	0,288
13	0,679	0,393	0,286
2	0,739	0,461	0,278
14	0,715	0,454	0,261
15	0,733	0,484	0,249
5	0,678	0,454	0,224
1	0,618	0,409	0,209
8	0,758	0,561	0,197
18	0,765	0,571	0,194
23	0,640	0,459	0,181
3	0,770	0,597	0,173
4	0,635	0,464	0,171
17	0,710	0,543	0,167
7	0,774	0,618	0,156
11	0,709	0,554	0,155
21	0,617	0,486	0,131
20	0,699	0,580	0,119
12	0,765	0,647	0,118
26	0,626	0,572	0,054
16	0,600	0,612	-0,012
Moy.	0,718	0,482	0,236
E.t.	0,073	0,109	0,141

Tableau 5 : classement des questions en fonction des écarts des certitudes moyennes

question du test CTS de mars 1998 (test sanctionnant), l'axe des abscisses reprend la moyenne des certitudes qui ont accompagné chaque réponse correcte et l'axe des ordonnées la moyenne des certitudes liées aux distracteurs. On visualise sous la diagonale, les questions dont la moyenne des certitudes liées aux réponses incorrectes (CMRIq) est moins élevée que la certitude moyenne de la réponse correcte (CMRCq). Plus le point représentant une question est proche du coin inférieur droit, plus l'utilisation des degrés de certitude dans le cadre de cette question est cohérente. A l'inverse, placés au-dessus de la diagonale, les points témoignent de fonctionnements incohérents. Les écarts entre les CMRCq et les CMRIq des questions 16 et 26 sont peu élevés : la question 16 se situe légèrement au-dessus de la diagonale et la 26 en est assez proche. Le classement des questions en fonction des écarts entre les CMRCq et CMRIq met en évidence en bas du tableau 5 les questions qui présentent un problème du point de vue de la cohérence métacognitive.

6. Comparaison des *r*.bis spectraux (*r*.bis S)

Le recours aux degrés de certitudes offre l'avantage de permettre le calcul d'un *r*.bis spectral (*r* .bis S) indépendant des scores globaux. Le *r*.bis S est calculé en corrélant les degrés de certitude (de 0 à 5) qui ont accompagné les réponses à une QCM avec les choix/rejets des propositions de cette QCM. Le *r*.bis-S est un indicateur de la qualité de la cohérence métacognitive d'un groupe d'étudiants (ou de testés) liée à chaque proposition d'une QCM. Il montre dans quelle mesure le choix de la proposition correcte est corrélé avec un degré de certitude élevé et inversement dans quelle mesure le choix des distracteurs est corrélé avec des degrés de certitude peu élevés (lorsque c'est le cas on obtient des corrélations négatives).

Les *r*.bis S ont été calculés pour chaque question du test CTS d'octobre 1998 (sanctionnant) à l'aide de SCANTEST dans le cadre d'une *turbo-analyse spectrale* avec un seuil de sélection des étudiants les plus réalistes à .81. L'exemple du tableau 6 concerne la première question pour laquelle la proposition n° 6 (SGI « aucune ») était la réponse correcte.

r.bis S	-0,085	-0,085	-0,150	-0,251	XXXX	-0,251	-0,252	-0,036	-0,085	-0,080
	10	17	13	13	14	12	19	17	18	19

Tableau 6 : exemple de r.bis S calculés pour une question – logiciel SCANTEST

SCANTEST offre également une représentation graphique qui synthétise l'information relative aux *r*.bis S pour l'ensemble des questions d'un test. Pour chaque question, l'axe des abscisses reprend la valeur du *r*.bis S lié à la réponse correcte et l'axe des ordonnées la valeur de la moyenne des *r*.bis S des propositions incorrectes. Pour une question, plus les réponses correctes sont accompagnées de degrés de certitude élevés et les réponses incorrectes de degrés de certitude faibles, plus la cohérence métacognitive du groupe de testés vis à vis de la question est élevée. Idéalement, les réponses correctes doivent recueillir des *r*.bis S positifs (colonne RC) et les distracteurs des *r*.bis S négatifs (la figure 8 reprend dans la colonne RI, la moyenne des *r*.bis des distracteurs). Une *zone de qualité* (coin inférieur droit) est mise en évidence sur le graphique, les diagonales de cette zone constituent des *niveaux de qualité*. Par exemple, sur le graphique du test CTS de mars 1998, les questions 6 et 9 se situent au *niveau 10*, la question 14 au *niveau 6*, etc. Les valeurs des niveaux de qualité sont reprises dans la colonne Niveaux de Qualité (NQ). Le niveau de qualité de la cohérence métacognitive des réponses des questions 16 et 26 est à 0 car ces questions se situent en dehors de la *zone de qualité*. Plus le niveau de qualité de la cohérence métacognitive des réponses d'une question est élevé, plus elle se situe dans un couloir de qualité proche du coin inférieur droit du graphique.

r.bis S

Qst	RC	RI	NQ
1	0,370	-0,074	**5**
2	0,463	-0,045	**6**
3	0,369	-0,151	**6**
4	0,323	-0,095	**5**
5	0,342	-0,152	**5**
6	0,604	-0,344	**10**
7	0,304	-0,136	**5**
8	0,308	-0,144	**5**
9	0,659	-0,270	**10**
10	0,487	-0,180	**7**
11	0,318	-0,128	**5**
12	0,213	-0,042	**3**
13	0,541	-0,215	**8**
14	0,469	-0,087	**6**
15	0,417	-0,161	**6**
16	-0,055	0,110	**0**
17	0,380	-0,086	**5**
18	0,391	-0,126	**6**
19	0,583	-0,208	**8**
20	0,195	-0,050	**3**
21	0,260	-0,085	**4**
22	0,632	-0,183	**9**
23	0,356	-0,106	**5**
24	0,566	-0,209	**8**
25	0,636	-0,217	**9**
26	0,036	0,107	**0**

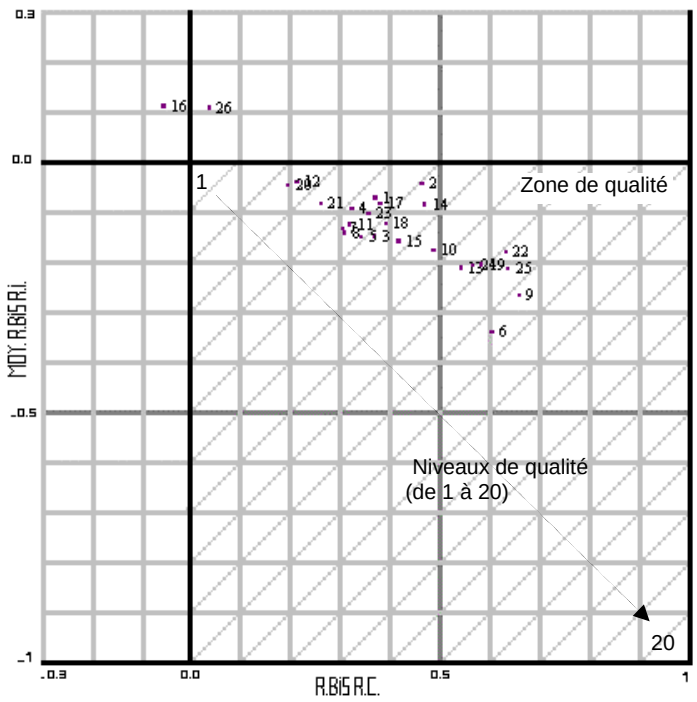

Figure 8 : r.bis S après turbo-analyse spectrale (t.81$_{80}$) du test CTS de mars 1998

Qst.	Niveaux de qualité des questions estimés sur base des			Alpha de Cronbach
	r.bis spectraux calculés à partir des résultats des étudiants réalistes (N=80)	r.bis classiques calculés à partir des résultats de tous les étudiants (N=247)	**Ecarts**	alpha si suppression de la question
8	5	5	**0**	0,720
21	4	4	**0**	0,726
23	5	5	**0**	0,722
26	0	0	**0**	0,736
1	5	6	**1**	0,714
3	6	7	**1**	0,709
4	5	4	**1**	0,724
5	5	6	**1**	0,708
13	8	7	**1**	0,713
17	5	6	**1**	0,714
18	6	5	**1**	0,715
2	6	4	**2**	0,718
7	5	7	**2**	0,714
15	6	4	**2**	0,723
20	3	5	**2**	0,724
9	10	7	**3**	0,718
10	7	4	**3**	0,725
11	5	8	**3**	0,711
19	8	5	**3**	0,714
22	9	6	**3**	0,709
25	9	6	**3**	0,715
6	10	6	**4**	0,719
12	3	7	**4**	0,712
14	6	2	**4**	0,725
16	0	5	**5**	0,722
24	8	2	**6**	**0,735**

Tableau 7 : écarts entre les niveaux de qualité r.bis S et r.bis classiques des questions et comparaison avec les coefficients alpha de Cronbach

Nous avons comparé les classements obtenus à partir des niveaux de qualité des *r*.bis spectraux issus de la *turbo-analyse* (calculés à partir des résultats des étudiants les plus réalistes) avec les *r*.bis « classique » (calculés sur l'ensemble des étudiants). La colonne « Ecarts » du tableau 7 reprend les écarts (en gras) entre les niveaux de qualité estimés à partir des *r*.bis spectraux et ceux estimés à partir des *r*.bis d'une analyse classique. Les écarts sont nuls pour les questions 8, 21, 23 et 1. Parmi ces

questions, la 26 se situe en dehors de la zone de qualité. En bas du tableau 7, là où les écarts sont grands entre les niveaux de qualité livrés par les deux analyses, apparaissent les questions 16 et 24. Pour ces questions, les analyses aboutissent à des résultats contradictoires dans la mesure où à un niveau plutôt bon dans un cas, correspond un niveau plutôt mauvais dans l'autre. Du point de vue de la cohérence interne[11] (mesurée à l'aide des *r*.bis classiques) les questions 16 et 26 posent des problèmes d'incohérence interne. Du point de vue de la cohérence métacognitive (*r*.bis S), la question 26 pose également un problème d'incohérence métacognitive. Lors du traitement nous avons procédé au calcul de l'alpha de Cronbach qui vaut 0,727 pour cette épreuve. La dernière colonne du tableau 7 montre les coefficients alpha qu'obtiendrait le test si la question était supprimée.

Les trois analyses concordent en ce qui concerne la question 26 qui obtient un niveau de qualité 0 du point de vue de la cohérence interne et de la cohérence métacognitive, de plus, si on la supprimait, elle amènerait la plus forte amélioration de l 'alpha de Cronbach : 0,736. Par contre, pour la question 24 elles divergent : l'alpha après suppression est de 0,735 (juste après la question 26), le niveau de qualité de cohérence interne est relativement bas (2) mais pas à 0, et, du point de vue de la cohérence métacognitive, le niveau est assez élevé (8). Les corrélations des résultats obtenus par les différentes analyses sont résumées dans le tableau 8. Nous avons inversé le signe de la corrélation lorsque l'alpha après suppression était impliqué dans la corrélation, car plus l'alpha après suppression de la question est élevé, moins la performance de la question est bonne. Nous avons aussi corrélé les *r*.bis S, *r*.bis et alpha avec le Rq (calculé sur les étudiants les plus réalistes). Les corrélations significatives apparaissent en gras accompagnées des seuils de probabilité. On observe des corrélations significatives élevées pour l'alpha avec le *r*.bis, ce qui nous paraît assez logique car l'alpha mesure aussi la cohérence interne, à ce propos, DE LANDSHEERE (1979) souligne : « *Alpha est, en fait, la moyenne de tous les coefficients de bipartition possibles pour un même test* ».

	r.bis S	*r*.bis	Alpha	Rq
r.bis S				
r.bis	0,30 (p=0,137)			
Alpha	0,29 (p=0,153)	**0,86** (p=0,000)		
Rq	**0,59** (p=0,001)	**0,54** (p=0,005)	0,28 (p=0,161)	

Tableau 8 : r des analyses de questions

La corrélation positive et significative de l'indice Rq (Réalisme par question) avec le *r*.bis S va dans le sens d'une logique de fonctionnement métacognitif selon laquelle la question qui ne favorise pas la cohérence métacognitive aura tendance à obtenir un score de Réalisme de la question (Rq) peu élevé, ce qui est le cas pour les questions 16 (NQ r.bis S = 0 et Rq = 0,575) et 26 (NQ r.bis S = 0 et Rq = 0,638).

La corrélation positive et significative du *r*.bis « classique » avec le Rq semble indiquer une laison entre, d'une part le fait qu'une question déclenche des réponses correctes chez les étudiants performants à l'ensemble du test et inversement déclenche des réponses incorrectes chez les étudiants moins performants (*r*.bis S) et, d'autre part le fait que la question induise chez les étudiants le choix de probabilités subjectives de réussite dont les valeurs centrales sont proches de leurs taux d'exactitudes (Rq). Si cette observation était confirmée par d'autres études, nous pourrions faire l'hypothèse d'un lien entre *cohérence interne* et *cohérence métacognitive*.

7. Comparaison du réalisme des étudiants lors du test CTS de mars 1998 avec leurs performances académiques lors des examens partiels de janvier 1998

Existe-t-il une relation entre le réalisme des étudiants au test CTS et leur niveau de performance académique ? Nous avons tenté de répondre à cette question en corrélant les scores obtenus aux partiels de janvier avec les scores à l'indice de réalisme calculé à partir des réponses au test CTS de mars 1998. Cette corrélation vaut **0,47** (*r* significatif à $p < 0,05$ et N = 200 après suppression des variables à valeur manquante), ce qui semble indiquer qu'une telle relation existe.

Lorsqu'on calcule le réalisme à partir des données des 17 questions dont les réponses attendues étaient des SGI, la corrélation s'améliore légèrement : elle vaut **0,50**. Par contre pour les 9 questions dont les réponses étaient des solutions « habituelles » (une des propositions dactylographiées) elle chute à **0,16**.

[11] Il y a cohérence interne pour une question lorsque les choix/rejets de la réponse correcte et des distracteurs sont corrélés (cohérents) avec les scores des étudiants à l'ensemble des questions du test.

8. Conclusions

Nous assistons aujourd'hui à une augmentation du nombre d'étudiants dans l'enseignement supérieur. Face aux grands groupes, beaucoup d'enseignants sont décidés à recourir à des dispositifs de correction d'examen automatisés. A l'Université de Liège, le SMART du CAFEIM tente de répondre à cette demande tout en veillant à garantir aux enseignants et aux étudiants une qualité docimologique satisfaisante. C'est la raison pour laquelle nous préconisons l'utilisation des SGI et des degrés de certitude lorsque les enseignants veulent recourir aux QCM. Lors d'une étude précédente, GILLES et LECLERCQ (1995) ont mis en évidence un cycle de réalisation des évaluations certificatives. Nous pensons que l'intégration de ce cycle dans une boucle de qualité ISO 9004-2 serait bénéfique à la fois pour les enseignants et pour les étudiants, destinataires du service, mais aussi pour l'institution universitaire qui pourrait se prévaloir auprès de ses futurs étudiants de cette démarche qualité. L'intégration du processus dans une boucle qualité nécessite une réflexion « assurance qualité » à chaque étape du cycle SMART. En ce qui concerne l'étape 6 « correction de l'examen », à notre connaissance, peu de chercheurs se sont penchés sur l'utilisation des informations métacognitives livrées par l'utilisation des degrés de certitude pour évaluer la qualité des questions.

En ce qui concerne l'exploitation des informations métacognitives en vue de mesurer les performances des étudiants, de nouvelles approches sont en train de voir le jour, notamment grâce aux travaux de JANS (1998), JANS et LECLERCQ (1998). Dans le cadre de cette étude, nous avons pu établir un lien entre les performances académiques lors d'examens et le réalisme des étudiants lors du test CTS de mars 1998. La corrélation de 0,47 est heureuse étant donné qu'il s'agit de personnes qui se destinent à la profession médicale, particulièrement sensible aux problèmes de réalisme. Au vu de la corrélation des performances académiques avec le réalisme aux QCM SGI (r = 0,50), nettement plus élevée que la corrélation des performances académiques avec le réalisme aux QCM habituels (r = 0,16), nous formulons l'hypothèse que la force de cette relation est liée au niveau taxonomique des questions posées. Si on se réfère à la taxonomie de BLOOM, les QCM habituelles relèvent le plus souvent du niveau connaissance, compréhension et application tandis que les QCM SGI relèvent de niveaux taxonomiques plus élevés tels que l'analyse et l'évaluation.

Pour évaluer qualité des QCM grâce aux informations métacognitives livrées par l'utilisation des degrés de certitude nous avons mis au point le logiciel SCANTEST. Ce programme nous permet de calculer et visualiser (1) les profils spectraux par question, (2) le réalisme par question (Rq), (3) les *r*.bis spectraux (*r*.bis S) en plus des *r*.bis classiques. Nous avons également tenté de clarifier les principes de base d'une *turbo-analyse spectrale* qui permet d'augmenter la fiabilité des analyses spectrales en nous fondant sur les résultats des étudiants les plus réalistes. Dans le cadre de cette étude, les analyses qui n'utilisent pas l'information métacognitive et qui sont centrées sur la cohérence interne (les *r*.bis classiques et l'alpha de Cronbach), montrent pour la question la plus problématique de l'épreuve (question 26) une concordance avec l'analyse spectrale centrée sur la cohérence métacognitive (*r*.bis S et Rq). Lorsqu'on se penche sur les corrélations des *r*.bis S, *r*.bis, alpha et Rq, on observe des corrélations positives significatives « intra-approches » : 0,86 pour la cohérence interne (*r*.bis *vs* alpha) et 0,59 pour la cohérence métacognitive (*r*.bis S *vs* Rq). Pour ce qui est des corrélations « inter-approches », une corrélation positive et significative est observée lorsqu'on compare les niveaux de qualité de l'analyse *r*.bis classique avec les scores de réalisme par question (Rq).

Le calcul des *r*.bis S et des Rq pour une question donnée ne prend pas en compte les autres questions du test. Cette indépendance des *r*.bis S et des Rq, indices de cohérence métacognitive d'une question, en font de précieux indicateurs de qualité dans la perspective d'une gestion de banque de questions. Dans cette perspective, les profils spectraux des questions trouveront aussi leur place, d'autant plus qu'on observe une remarquable stabilité des profils pour les questions du test d'entraînement soumis à deux groupes d'étudiants différents (voir figure 3 et tableau 3). Enfin, les représentations des corrélations des profils spectraux avec les profils théoriques sur des ingénogrammes (voir figure 4) devraient également faciliter les classements et les tris dans le cadre d'une banque de questions.

9. Bibliographie

DE LANDSHEERE, G. (1979). *Dictionnaire de l'évaluation et de la recherche en éducation*, Paris : Presses universitaires de France.

GILLES, J-L. (1996). Utilisation des degrés de certitude et normes de réalisme en situation d'examen et d'auto-estimation à la Faculté de Psychologie et des Sciences de l'Education de l'Université de Liège, *Actes du Colloque de l'ADMEE-EUROPE « Dix années de travaux de en évaluation »* - septembre 1996 - Université Pierre Mendes France à Grenoble, à paraître.

GILLES, J.-L. (1998). Apports des mesures métacognitives lors d'un test de sélection portant sur la compréhension d'un article scientifique en 1ère candidature de la Faculté de Médecine, *Actes du 12ème Colloque de l'ADMEE, Mons, septembre 1998*, à paraître aux éditions De Boeck.

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

GILLES, J.-L. (1997). Impact de deux entraînements à l'utilisation des degrés de certitude chez les étudiants de 1ère candidature de la Faculté de Psychologie et des Sciences de l'Education de l'ULg, *Actes du 15ème Colloque AIPU*, Liège : Affaires Académique de l'Université de Liège.

JANS, V. (1998). L'auto-évaluation de performances simples et complexes par des étudiants universitaires : description et résultats d'une expérience, *Actes du 12ème Colloque de l'ADMEE, Mons*, septembre 1998, à paraître aux éditions De Boeck.

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

LECLERCQ, D. & al. (1993). Validity, reliability and acuity of self-assessment in educational testing. NATO ASI Series, *Item Banking: Interactive Testing and Self Assessment*, Berlin: Springer Verlag, 1993, Vol. 112, pp. 114-131.

LECLERCQ, D. (1998). Mesurer l'effet de l'apprentissage à l'aide de l'analyse spectrale des performances *Actes du 12ème Colloque de l'ADMEE, Mons*, septembre 1998, à paraître aux éditions De Boeck.

MELON, S. et GILLES, J.-L. (1998).

SHUFFORD, E. & al. (1966). *Admissible Probability Measurement Procedures*, Psychometrika, 31, 125-145.