Apports des mesures métacognitives lors d'un test de compréhension d'un article scientifique

Jean-Luc Gilles
Université de Liège
Belgique

Introduction

Depuis l'année académique 1994-1995, le Centre d'auto-formation et d'évaluation interactives Multimédias (CAFEIM) de la faculté de psychologie et des sciences de l'éducation (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* (formulaires spéciaux pour la lecture optique de marques) ;
- le testing informatisé interactif où les réponses sont fournies à l'aide des ordinateurs du CAFEIM (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 l'année académique 1998-1999) qui permettent de tester 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 3^{e} candidature dès l'an 2000. Un des tests de sélection, réalisé en collaboration 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. Le test de compréhension d'un article scientifique

1.1 Description

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 1re candidature de cette faculté. Les questions portent sur la compréhension d'un texte tiré d'une revue de vulgarisation scientifique. L'épreuve reprise dans cette étude eut lieu en mars 1998 et fut précédée par un entraînement qui débuta dès octobre 1997. Les

questions étaient à choix multiple avec Solutions Générales Implicites (SGI)[1]. 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. Le temps imparti pour le test était de 90 minutes. Une première simulation en grandeur réelle eut lieu en mars 1997 avec un groupe d'étudiants volontaires (N=130) de 1re candidature de l'année académique 1996-1998. Le test utilisé à cette occasion fut ensuite proposé dans le cadre de la procédure d'entraînement aux étudiants de l'année académique 1997-1998, ceux pour qui le test final serait sanctionnant.

1.2 Modalités d'utilisation des degrés de certitude

Une série de règles méthodologiques sont respectées lorsqu'on utilise les degrés de certitude : (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. Le tableau 1 présente le barème des tarifs utilisé (Leclercq, 1983, 1993).
De nombreuses façons d'exprimer le degré de certitude ont été décrites (voir Leclercq, 1993, pp. 114-131). Seules celles qui respectent les 4 règles énoncées plus haut sont considérées comme « *Admissible Probability Measurement Procedures* » par Shufford & al., (1966).

Tableau 1 : Barème des tarifs liés aux degrés de certitude de D. Leclercq

Si vous considérez que votre réponse a une probabilité d'être correcte comprise entre	Ecrivez	Vous obtiendrez 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

2. SCANTEST, un logiciel pour mesurer la qualité des questions en utilisant l'information livrée par les degrés de certitude

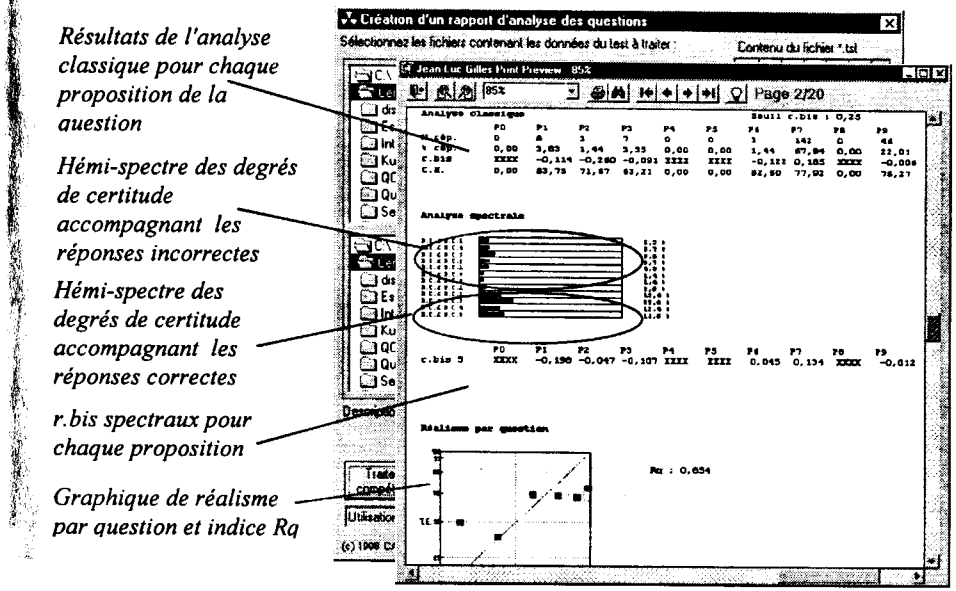
De nouvelles méthodes permettant d'évaluer la qualité des questions en ayant recours à l'information livrée par les degrés de certitude ont été développées et mise en oeuvre à l'aide d'un logiciel intitulé SCANTEST (voir figure 1). Ce programme offre :

- une analyse classique des propositions de chaque QCM : nombre d'étudiants (N rép.) qui ont choisi la proposition, pourcentage (% rép.), corrélation bisériale de point (*r*.bis) et certitude moyenne (C.M.);

[1] 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 (contrevérité dans l'énoncé *à dénoncer en priorité !*).

- une analyse spectrale avec 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) ;
- 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 de réalisme adaptée par Gilles (1997).

Figure 1 : Exemple de rapport de traitement pour une question - logiciel SCANTEST



Des synthèses sont également fournies :

- un récapitulatif des certitudes moyennes pour les réponses correctes et incorrectes accompagné d'un graphique de cohérence métacognitive;
- 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.

Les profils spectraux, le Réalisme par question (Rq), les Certitudes Moyennes pour les Réponses Correctes (CMRCq) et Incorrectes (CMRIq), les *r*.bis spectraux (*r*.bis S) et leurs niveaux de qualité (NQ), sont détaillés dans les paragraphes qui suivent.

3. Analyse spectrale des questions

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

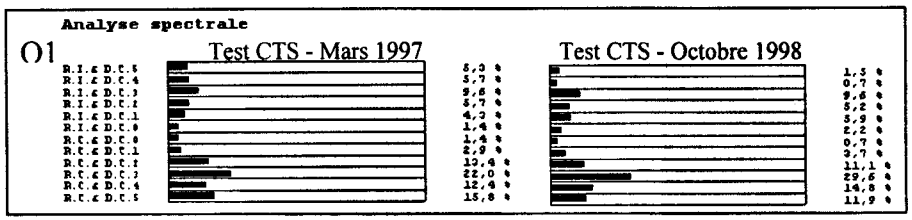
Rappelons que le test d'entraînement fut soumis une première fois en mars 1997 à un premier groupe d'étudiants de 1re candidature (N=199) et une seconde fois en octobre 1997 à un second groupe de 1re candidature (de l'année académique suivante, N=125). La figure 2 montre les profils spectraux de la question 1 remarquablement similaires pour les deux groupes. Faute de place, nous n'avons pu reproduire ici les 30 autres profils des questions. Tous sont très proches, sauf les profils spectraux relatifs à la question 12.

Le tableau 2 reprend les corrélations des pourcentages liés aux profils spectraux des 16 questions du test d'entraînement soumises en 1996-1997 avec les pourcentages des profils spectraux de 1997-1998. On remarque des corrélations très élevées pour la plupart des questions, sauf pour la question 12 (r = 0,52). La moyenne des corrélations des 16 questions est élevée (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 2 : r CTS entraînement 1997 vs.1998

Figure 2 : Comparaison des profils spectraux de la question 1 du test CTS d'entraînement (1996-1997 et 1997-1998)



3.2 Profils spectraux types

Le profil spectral de la répartition des degrés de certitude peut être corrélé avec des profils types :

- 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 utilisations 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 types

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.

Tableau 3 : Corrélation du profil spectral de la question 1 du test CTS d'entraînement avec les profils spectraux théoriques

Question 1		Profils théoriques FMR	DMR	FMNR	DMNR	HLRI	HLRC	HGRI	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

Dans le cas de la question 1 posée en mars 1997, on constate des corrélations positives (voir la dernière ligne du tableau 3) 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[2], 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 3), 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

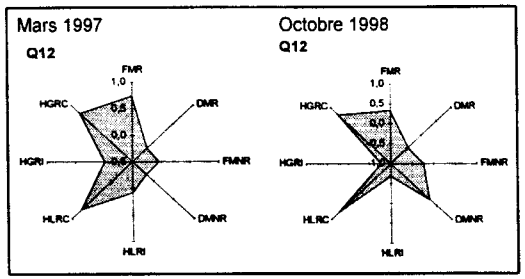


Figure 3 : profils Q12-03/97 et Q12-10/98

[2] 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). Les points d'intersection des lignes qui relient les axes représentent les valeurs des r.

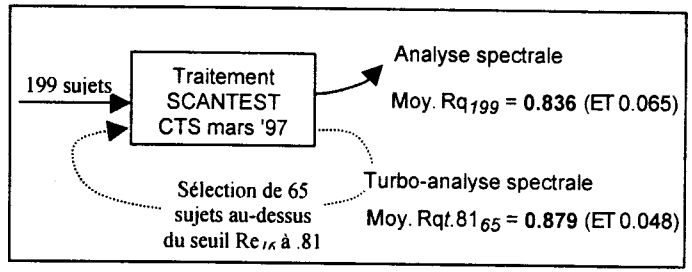
d'approche typologique pourrait se révéler particulièrement utile lorsqu'il s'agit de sélectionner des questions en vue de créer des tests équivalents.

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) pour ensuite analyser leurs seules réponses à l'aide du 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) 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 considéré 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).



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. Pour le test CTS d'entraînement qui s'est déroulé en mars 1997, nous constatons une amélioration de 0,043 (0,879 – 0,836) de la moyenne des Rq après la *turbo-analyse spectrale*. L'amélioration après la *turbo-analyse spectrale* se visualise sur la plupart graphiques de réalisme par question (dans 12 cas sur 16) par un 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.

5. Analyse 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 positif généralement obtenu pour la réponse correcte montre dans quelle mesure le choix d'une proposition correcte est corrélé avec un degré de certitude élevé. Lorsqu'on calcule le *r*.bis S d'un distracteur sur l'ensemble des sujets, la corrélation des choix/rejet (1/0) avec les degrés de certitude (0 à 5) est en principe négative ou peu élevée et montre dans quelle mesure les choix du distracteur ont été accompagnés de degrés de certitude faibles et les rejets de degrés de certitude plus élevés. Idéalement, pour établir le *r*.bis S d'un distracteur on ne devrait prendre en compte que (1) les réponses des étudiants qui ont choisi ce distracteur et (2) les réponses des étudiants qui ont répondu correctement. En effet, les « 0 » des rejets du distracteur désignent deux types de situations : celles où la réponse correcte a été sélectionnée (on s'attend alors à un degré de certitude plus élevé) ou celles où un autre distracteur a été choisi (on s'attend alors à un degré de certitude moins élevé). En opérant une sélection des réponses des étudiants pour calculer le *r*.bis S on obtiendrait des valeurs plus proches de -1 que celles qui figurent dans les exemples ci-après (SCANTEST calcule actuellement les *r*.bis des distracteurs en prenant toutes les réponses).

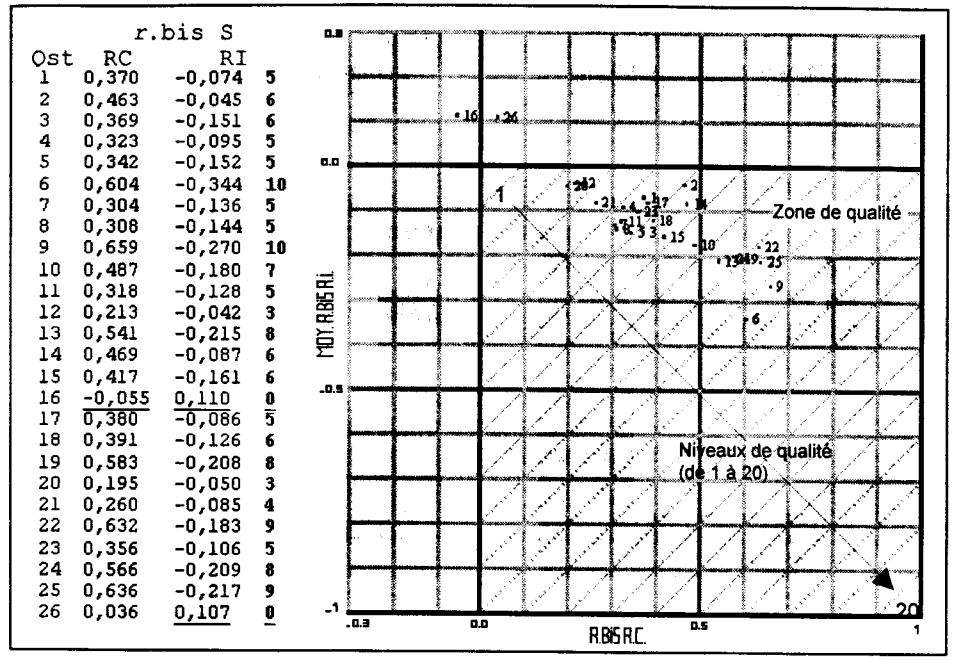
Les *r*.bis S ont été calculés pour chaque question du test CTS d'octobre 1998 (sanctionnant) à l'aide de SCANTEST dans le cadre d'une *turbo-analyse spectrale* avec un seuil de sélection des étudiants les plus réalistes à .81. L'exemple du tableau 4 concerne la première question pour laquelle la proposition n° 6 (SGI « aucune ») était la réponse correcte.

Tableau 4 : exemple de r.bis S calculés pour une question – logiciel SCANTEST

	P0	P1	P2	P3	P4	P5	P6	P7	P8	P9
r.bis S	-0,290	-0,249	0,027	-0,125	-0,142	XXXX	0,370	-0,145	-0,142	0,083

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. Idéalement, les réponses correctes doivent recueillir des *r*.bis S positifs (colonne RC) et les distracteurs des *r*.bis S négatifs (la figure 4 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 des *r*.bis S 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 des *r*.bis S d'une question est élevé, plus elle se situe dans un couloir de qualité proche du coin inférieur droit du graphique.

Figure 4 : r.bis S après turbo-analyse spectrale (t.81$_{80}$) du test CTS de mars 1998



Nous avons comparé les niveaux de qualité de *r*.bis S spectraux issus de la *turbo-analyse* (calculés à partir des résultats des étudiants les plus réalistes) avec les niveaux de qualité obtenus à l'aide des *r*.bis « classiques » (calculés sur l'ensemble des étudiants). 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. 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 pour un type de r.bis, correspond un niveau plutôt mauvais pour l'autre type. Du point de vue des *r*.bis classiques, les questions 16 et 26 posent des problèmes. Du point de vue des *r*.bis S la question 26 pose également un problème. Lors du traitement nous avons procédé au calcul de l'alpha de Cronbach qui vaut 0,727 pour cette épreuve. Nous avons aussi calculé les coefficients alpha qu'obtiendrait le test si ces questions étaient supprimées. Les trois analyses concordent en ce qui concerne la question 26 qui obtient pour les *r*.bis et *r*.bis S un niveau de qualité 0, 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 du *r*.bis est relativement bas (2) mais pas à 0, et, du point de vue du *r*.bis S, le niveau est assez élevé (8). Les corrélations des résultats obtenus par les différentes analyses sont résumées dans le tableau 5. 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 question participe à la cohérence interne du test. Nous avons aussi corrélé les *r*.bis S, *r*.bis et alpha avec le Rq (calculé comme le *r*.bis S 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 ces deux coefficients sont indicateurs de 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* ». La corrélation positive et significative de l'indice Rq (Réalisme par question) avec le *r*.bis S semble indiquer dans le cadre de ce test qu'un Réalisme de la question (Rq) faible s'accompagne d'un Niveau de Qualité (NQ) du *r*.bis S peu élevé et inversement. La corrélation positive et significative du *r*.bis « classique » avec le Rq montre une liaison entre le fait que les questions discriminent bien les étudiants performants à l'ensemble du test et le fait que la question induise le choix de probabilités subjectives de réussite dont les valeurs centrales sont proches de leurs taux d'exactitudes (Rq).

	r.bis S	*r*.bis	Alpha
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 5 : r des analyses de questions

6. 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'adoption d'une démarche qualité selon les normes de la série ISO 9001 pour la réalisation des évaluations serait bénéfique à la fois pour les enseignants et pour les apprenants, mais aussi pour l'institution universitaire qui pourrait se prévaloir auprès de ses futurs étudiants de cette approche qualité. L'intégration du cycle dans une boucle qualité nécessite une réflexion « assurance qualité » à chaque étape du processus SMART. En ce qui concerne l'étape « 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.

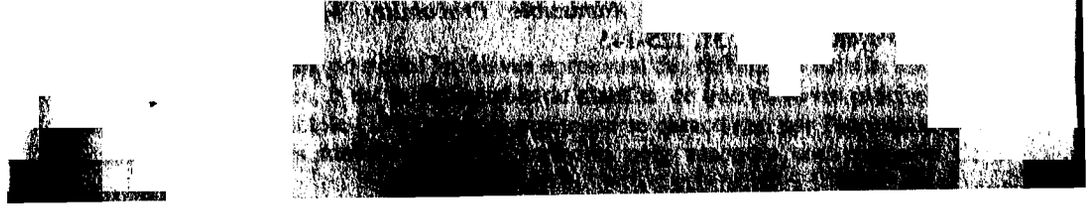
Pour évaluer qualité des QCM grâce aux informations métacognitives livrées par l'utilisation des degrés de certitude nous avons mis au point le logiciel SCANTEST. Ce programme nous permet de calculer et visualiser (1) les profils spectraux par question, (2) le réalisme par question (Rq), (3) les *r*.bis 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 liée à l'utilisation des degrés de certitude 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 (*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 pour les indicateurs de cohérence interne (*r*.bis *vs* alpha = 0,86) et pour les indicateurs de l'analyse spectrale (*r*.bis S *vs* Rq = 0,59). Une corrélation positive et significative est observée lorsqu'on compare les niveaux de qualité des *r*.bis classiques 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, en font de précieux indicateurs de qualité pour la gestion d'une banque de questions. Dans cette perspective, les profils spectraux des questions trouveront aussi leur place (nous observons une remarquable stabilité des profils pour les questions du test d'entraînement soumis à deux groupes d'étudiants différents - voir figure 2 et tableau 2). Enfin, les représentations des corrélations des profils spectraux avec les profils théoriques sur des ingénogrammes (voir figure 3) devraient également faciliter les classements et les tris dans le cadre de la gestion d'une banque de questions.

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

Jean-Luc Gilles

CAFEIM-FAPSE
Université de Liège
Boulevard du Rectorat, 5 – Bât B32
4000 Liège (Sart Tilman)
Belgique
E-Mail : jl.gilles@ulg.ac.be



Bibliographie

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

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. (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 recherche en évaluation »* - septembre 1996 - Université Pierre Mendes France à Grenoble, à paraître.

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 15e Colloque AIPU,* Liège : Affaires Académique de l'Université de Liège.

Gilles, J.-L. (1998)., Mise en oeuvre de tests formatifs à l'aide de l'internet dans les enseignements de la Faculté de psychologie et des sciences de l'éducation – ULg : 1er bilan et perspectives. *Actes du 12e Colloque de l'ADMEE, Mons*, septembre 1998, à paraître aux presses de l'université de Mons-Hainaut.

Jans, V. (1998). L'autoévaluation de performances simples et complexes par des étudiants universitaires : description et résultats d'une expérience, *Actes du 12e Colloque de l'ADMEE, Mons*, septembre 1998, à paraître.

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 12e Colloque de l'ADMEE, Mons*, septembre 1998, à paraître aux éditions De Boeck.

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